APPLYING SEMANTIC TECHNOLOGIES AND ARTIFICIAL INTELLIGENCE TO ECO-INFORMATIC MODELLING OF CORAL REEF SYSTEMS

Thesis submitted by

Trina Sharlene Myers

November 2009

for the Degree of Doctor of Philosophy

James Cook University

Supervisor:

Professor Ian Atkinson

Statement of Access

I the under-signed, the author of this work, understand that James Cook University will make this thesis available for use within the University Library and, via the Australian Digital Thesis network, for use elsewhere.

I understand that, as an unpublished work, a thesis has significant protection under the Copyright Act and I do not wish to place any restriction on access to this thesis.

_____ _____

Signature Date

Declaration

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

_____ _____

Signature Date

Electronic Copy

I the under-signed, the author of this work, declare that the electronic copy of this thesis provided by James Cook University Library is an accurate copy of the print thesis submitted, within the limits of the technology available.

_____ _____

Signature Date

Statement on the Contribution of Others

The research described and presented in this thesis was undertaken by the author under supervision by Professor Ian Atkinson and Professor Bill Lavery, both of whom provided editorial and academic advice.

For financial support, I thank Professor Marimuthu Palaniswami of the ARC research network on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), Department of Electrical and Electronic Engineering from the University of Melbourne for the guidance, support and scholarship in the later stages of the research. I am also grateful to the James Cook University Graduate Research School for an in-kind scholarship and the Faculty of Science and Engineering for two internal Research Grants. I am also appreciative of the travel allowances given by the (then) School of Maths, Physics and Information Technology (JCU) and the DEST funded ARCHER project that made possible the publication and subsequent presentations of this research at both national and international conferences.

I thank Dr. Rosemary Dunn with deep gratitude for her infectious love of the written word. Her editorial advice and tutorage in literature and writing skills exponentially increased the standard of this thesis.

I would like to thank Dr. Ron Johnstone, Dr. Glen Homes and Jeff Maynard for their helpful discussions and expertise in the marine domain.

Acknowledgments

Firstly I would like to thank my husband Michael for being my biggest fan. The journey has been both exciting and arduous and throughout it Michael has been there as a powerful source of motivation and strength with his eternally positive attitude. I would like to thank my parents, Dr. Lance and Beth Myers, for also being incredibly supportive. Their belief in my ability and potential has always been a strong influence in my life and I know, because of this belief, every challenge I undertake will be successfully accomplished.

I owe deep gratitude to my primary supervisor Professor Ian Atkinson who has been an amazing mentor. Ian has never wavered in his belief in me and my abilities, even in times when I questioned it myself. Ian's support and guidance throughout the journey was superb, he grounds a student and clears their mind in his unique way so the scope is always visible and the goals are always attainable. I would like to thank my secondary supervisor Professor Bill Lavery for his faith in me and his support throughout the journey.

Finally I would like to thank all of those that contributed to this thesis through sheer faith, collaboration, discussion, and/or feedback and with special thanks to: David Browning, Louise Dowling, Dr. Ickjai Lee, Dianna Madden, Associate Professor Richard Monypenny, Nigel Sim and Dr. Jarrod Trevathan.

List of Publications

Myers, T., Atkinson, I. & Johnstone, R. 2010, 'Semantically enabling the SEMAT project: extending marine sensor networks for decision support and hypothesis testing (accepted)', *3rd International Workshop on Ontology Alignment and Visualization (OnAV 10)*, Krakow, Poland, 15 - 18 February, IEEE.

Myers, T. S., Atkinson, I. M. & Johnstone, R. 2009, 'Supporting coral reef ecosystems research through modelling a re-usable ontology framework', *Journal of Applied Artificial Intelligence*, vol. 90, no. 24**,** pp. (in press).

Myers, T. S., Atkinson, I. & Johnstone, R. 2008, 'Supporting coral reef ecosystems research through modelling re-usable ontologies', *Proceedings from the Knowledge Representation Ontology Workshop (KROW 2008)*, Sydney, Australia, 17 September, ACS, pp. 51-59.

Myers, T. S. & Atkinson, I. M. 2008, 'The Semantic Reef: A hypothesis-based, eco-informatics platform to support automated knowledge discovery for remotely monitored reef systems.', *Proceedings of the 11th International Coral Reef Symposium (ICRS 08)*, Ft. Lauderdale, FL, USA, 7-11 July.

Myers, T. S., Atkinson, I. M. & Maynard, J. 2007, 'The Semantic Reef: An eco-informatics approach for modelling coral bleaching within the Great Barrier Reef', *Environmental Research Event (ERE 07)* Cairns, Australia, Environmental Research Event Organising Committee.

Myers, T. S., Atkinson, I. M. & Lavery, W. J. 2007, 'The Semantic Reef: Managing complex knowledge to predict coral bleaching on the Great Barrier Reef', *Proceedings of the fifth Australasian symposium on ACSW frontiers*, ACS, Ballarat, Australia, vol 68, pp. 59-67.

In loving memory of my beautiful little angel

Tameka

Abstract

A “data deluge” is overwhelming many areas of research. Massive amounts of scientific data are being produced that cannot be effectively processed. Remote environmental monitoring (including sensor networks) is being rapidly developed and adopted for collecting real-time data across widely distributed locations. As the volume of raw data increases, it is envisaged that bottlenecks will develop in the data analysis phase of research workflows, because data processing and synthesis procedures still generally involve manual manipulation.

Despite the exponential growth in data and the consequential challenges in data management, current e-Research communities are exploring solutions to the “data deluge”. E-Research is the amalgamation of research techniques, data and people with Information Communication Technologies (ICT) to enhance research capabilities. Recent research efforts by the Semantic Web and Knowledge Representation (KR) domains focus on the development of automated data synthesis technologies. A key component in these solutions is the semantic technologies. Semantic technologies involve methods to add contextual information to data through ontologies so logic systems can be applied by the computer to enable automated inference. An ontology explicitly describes concepts in “computer-understandable” terms which allows for automated reasoning and intelligent decision-making by the machine. Automated data analysis and knowledge discovery is desirable because the manual manipulation of data processing and synthesis requires human intervention which will become increasingly more difficult to sustain as the data deluge grows.

This dissertation introduces the Semantic Reef project which is an eco-informatics software architecture designed to alleviate data management problems within marine research. The intention was to develop an automated data processing, problem-solving and knowledge discovery system within the scope of e-Research, which will assist in developing our understanding and management of coral reef ecosystems. The Semantic Reef project employs e-Research approaches including semantic technologies and scientific workflows, which together create a platform designed to evaluate complex hypothesis queries and/or provide alerting for unusual events (e.g., coral spawning or bleaching).

The Semantic Reef project was built as a KR platform, so researchers can combine disjoint data from different sources into a single Knowledge Base (KB) to pose questions of the data. Scientific workflows access and retrieve remote sensor data and/or data available via the Web to

populate the KB. The KB consists of a hierarchy of reusable and usable ontologies that together generically model a coral reef ecosystem in a "computer-understandable" form. The ontologies range from informal through to formal and, when coupled to datasets, derive inferences from data to "ask" the KB questions for semantic correlation, synthesis and analysis. The ontology design leverages the scalable and autonomic characteristics of semantic technologies such as modularity, reuse and the ability to link latent connections in data through complex logic systems.

The overall goal of the Semantic Reef project was to enable marine researchers to pose hypotheses about environmental data gathered from *in situ* observations, and to explore phenomena such as climate change effects on an ecosystem rather than on one component at a time. Currently, in marine research, there has been an explosive increase in the number of questions posed about climate change effects; for example, questions about the origins of phenomena such as coral bleaching on coral reef ecosystems. To be answered, these questions need to be able to assess the cumulative combination of ecological factors and stressors that contribute to the tipping point from a healthy coral to stressed coral due to coral bleaching. The marine biology domain has an urgent need for more efficient investigation of the disparate data streams and data sources. The Semantic Reef project, which incorporates the new hypothesis-driven research tools and problem-solving methods, is designed as a proof of concept to resolve this need.

The Semantic Reef system has the capacity to pose hypotheses and automate inferences of the available data. The system's design supports flexibility in theoretic hypothesis design because the researcher is not required to predetermine the exact hypothesis prior to gathering data for import to the KB. Rather, the questions can be as flexible as the researcher requires, and they may evolve as new data becomes available or as ideas grow and/or epiphanies emerge. Then, once phenomena in the data are disclosed through semantic inference, *in situ* observations can be performed to confirm or negate the theory. The Semantic Reef tool offers marine researchers this flexibility in hypothesis modelling to theorise about a range of scientific conundrums such as the cumulative causal factors that contribute to coral bleaching.

This study is the first known example of Semantic Web technologies and scientific workflows combined to integrate data, with the purpose of posing observational hypotheses or inferring alerts in the coral reef domain. As a proof of concept, the Semantic Reef system offers a different approach to the development and execution of observational hypotheses on coral reefs. The system offers adaptability when applying hypotheses and questions of data, specifically in scenarios where the hypothesis is not apparent prior to data collection efforts. The Semantic Reef

system cannot overcome the data deluge, but it offers a unique approach to the discovery of new phenomena that, through automation, can alleviate the problems associated with the data analysis phase.

Table of Contents

List of Tables

List of Figures

Glossary of Acronyms

ACRONYM	MEANING
ABS	Australian Bureau of Statistics
AIMS	Australian Institute of Marine Science
API	Application Program Interface
AVHRR	Advanced Very High Resolution Radiometer
AWS	Automatic Weather Station
BOM	Australian Bureau of Meteorology
CC	Creative Commons
CHAMP	Coral Health and Monitoring Program
CRC	Cooperative Research Centre
CREON	Coral Reef Environmental Observatory Network
CSIRO	Commonwealth Scientific and Industrial Research Organisation
CWA	Closed World Assumption
DAML + OIL	Darpa Agent Markup Language plus the European Ontology Interchange Language
DHD	Degree Heating Day
DIG	Description Logic Implementation Group
DL	Description Logics
DLP	Description Logic Programming
DOGMA	Developing Ontology-Grounded Methods and Applications
FOL	First Order Logic
GBIF	Global Biodiversity Information Facility
GBR	Great Barrier Reef

ACRONYM	MEANING
GBRMPA	Great Barrier Reef Marine Park Authority
GBROOS	Great Barrier Reef Ocean Observing System
GEON	GEOscience Network
GIS	Geographic Information System
GLEON	Global Lake Ecological Observatory Network
HCI	Human Computer Interface
HPC	High Performance Computing
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
ICON	Integrated Coral Observing Network
IMOS	Integrated Marine Observing System
IPCC	Inter-governmental Panel on Climate Change
ICT	Information Communication Technologies
ITIS	Interagency Taxonomic Information System
JCU	James Cook University
JISC	Joint Information Systems Committee
JRE	Java Runtime Environment
JVM	Java Virtual Machine
KB	Knowledge Base
KR	Knowledge Representation
LHC	Large Hadrons Collider
LMSM	Local Mean Summer Maximum

ACRONYM	MEANING
LMST	Long-term Mean Sea Surface Temperature
LTMP	Long Term Monitoring Program
MMI	Marine Metadata Interoperability
MMM	Maximum Monthly Mean
MPL	Mozilla Public License
NAF	Negation as Failure
NEON	National Ecological Observatory Network
NEPTUNE	North-East Pacific Time-series Undersea Networked Experiments
NESDIS	National Environmental Satellite, Data and Information Service
NOAA	National Oceanic and Atmospheric Administration
NSF	National Science Foundation
OGC	Open Geospatial Consortium
OGSA	Open Grid Services Architecture
ONC	Ocean Networks Canada
OSG	Open Science Grid
OWA	Open World Assumption
OWL	Web Ontology Language
PAR	Photosynthetically Active Radiation
PROWL	Probabilistic Web Ontology Language (OWL)
RDF	Resource Description Framework
RDFS	RDF Schema
RIF	Rule Interchange Format

ACRONYM	MEANING
SBA	Satellite Bleach Alert
SEEK	Science Environment for Ecological Knowledge
SME	Subject Matter Expert
SIOC	Semantically-Interlinked Online Communities
SPARQL	SPARQL Protocol and RDF Query Language
SQWRL	Semantic Query-Enhanced Web Rule Language
SST	Sea Surface Temperature
SST+	SST anomaly
SSW	Semantic Sensor Web
SW	Semantic Web
SWEET	Semantic Web for Earth and Environmental Terminology
SWRL	Semantic Web Rules Language
uBio	Universal Biological Indexer and Organiser
UNA	Unique Name Assumption
URI	Unified Resource Identifiers
URL	Uniform Resource Locator
URN	Uniform Resource Name
VO	Virtual Organisations
VRE	Virtual Research Environment
W3C	World Wide Web Consortium
WWW	World Wide Web
XML	eXtensible Markup Language

Chapter One

Introducing the Semantic Reef Project

1.1. Chapter Synopsis

A "data deluge" is upon us! Massive amounts of digital data are being produced that cannot be effectively processed. Remote environmental monitoring (including sensor networks) for collecting real-time data across widely distributed locations is rapidly being developed and adopted. As the volume of raw data increases, it is envisaged that bottlenecks will develop in the data analysis phase of research workflows. Currently data processing procedures still generally involve manual manipulation, which will eventually become unfeasible to manage as the data collected exponentially grows (Hey and Trefethen 2003a).

Despite the seemingly overwhelming data deluge, recent research efforts by the Semantic Web and Knowledge Representation (KR) communities are exploring solutions to the "data deluge" problem (Goble et al. 2006; Hall et al. 2009). These efforts focus on the development of automated data synthesis technologies and use-case implementations. A key factor in this solution is the use of semantic technologies. Semantic technologies involve methods that add contextual information to data through ontologies, which makes the data computer-understandable, and therefore, automatically computer-processable or reasonable. This is desirable because the human effort required to interface with massive datasets is reduced by automating data processing and knowledge discovery.

This dissertation develops the Semantic Reef project. The Semantic Reef project is a software architecture designed to alleviate the problems in data management within very specific areas of marine science. The intention was to develop an automated data processing, problem-solving and knowledge discovery system that will assist in developing the understanding and management of coral reef ecosystems. The e-Research approaches employed by the Semantic Reef project include semantic technologies and scientific workflows, which together create a platform designed to evaluate complex hypothesis[1] enquiries and/or provide alerting for unusual events (e.g.,

[1] The term "hypothesis" in the context of this thesis means a proposed explanation for an observable phenomenon.

coral spawning or bleaching). The work outlined in this thesis is the first known attempt at providing data analysis and data processing solutions, with applied semantic technologies and scientific workflows, for observational hypothesis-driven research of coral reef ecosystems.

This chapter begins with an overview of e-Research and the problems found in modern data intensive research, specifically, the growth in data and the benefits of data integration. Following this, data management issues faced in the marine sciences, particularly the processing and analysis phase, is discussed. Then a description of the technologies that are being developed to alleviate the pressures from the growth of data and information on the Web and in research. In conclusion, the research objectives, the design approach of the Semantic Reef architecture, and the contributions and limitations, is presented.

1.2. E-Research

E-Research is the amalgamation of research techniques, data and people with Information Communication Technologies (ICT) to enhance research capabilities (Figure 1.1). Scientific progress increasingly depends on the sharing of resources, ideas, know-how and results. Some of

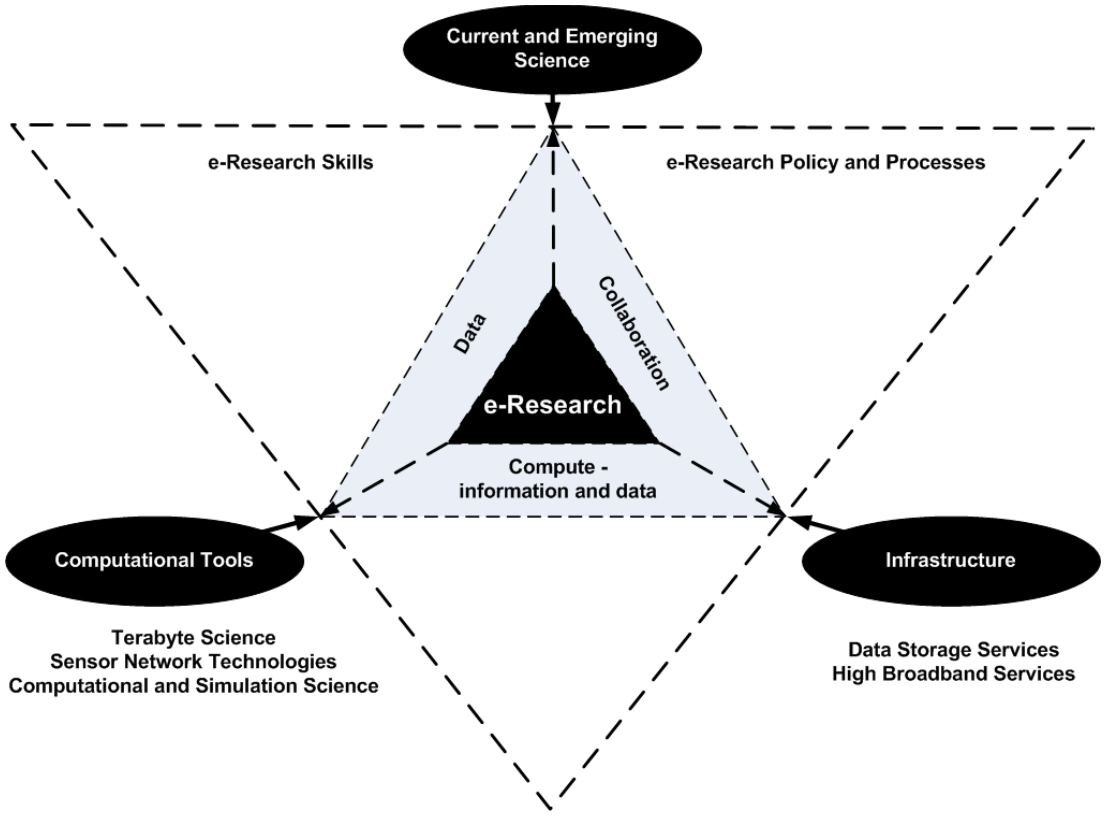


Figure 1.1. – e-Research, adapted from (Taylor et al. 2008)

the recent significant advances in science have been achieved through sharing complex interdisciplinary skills, data and analysis (Hendler 2003; Hey and Trefethen 2003b; Thorpe 2009). Connections between disconnected ideas, domains, people and data can clearly contribute to the creation of new knowledge. Additionally, the reuse of this knowledge, in ways not anticipated in its original creation, can subsequently be instrumental in the development of further knowledge (Goble et al. 2006).

To parallel the needs of researchers and thier quest for new knowledge, modern research tools and techniques are evolving. Researchers require access to specialised search engines, data mining tools, and data visualisation tools so asking questions and understanding the answers is abridged (Gil et al. 2007). Accordingly, to search, access, move, manipulate, and mine data stored in vast distributed digital repositories or discreet data silos, automatically, is a central requirement of the new generation of research software (Goble et al. 2006). The rearrangement and juxtaposition of inter-disciplinary data in interesting, efficient and exploratory ways often requires the application of high-throughput and/or High Performance Computing (HPC) and automated data integration methods. Further, the evolving requirements in data processing capabilities are in part motivated by the continual growth of scientific data acquisition (the data deluge) (Hey and Trefethen 2003a).

1.3. The Data Deluge

The exponential growth in accumulated scientific data is a key driver of the development efforts for new scientific tools to automate data integration, processing and analysis. The emergence of network connected and specialised data collection instrumentation, from large-scale synchrotrons to networks of tiny-remote sensors, has necessitated the efficient processing, organising, and creation of useful information from the mass of new data (Hey and Trefethen 2003a). The data deluge is a result of the growing influx of data and some of the main contributing sources are:

- The raw data outputs of scientific instruments;
- The metadata that is attached to the raw data (e.g., provenance information);
- The *in silico* simulations and modelling carried out on the data, which themselves become data sources for use in future studies; and
- The documents, data and information available on the World Wide Web (WWW, or Web) continue to grow at an overwhelmingly exponential rate (Adamic and Huberman

2002). Much of the "deep Web" (i.e., data repositories, data silos and server side databases) offers rich sources of information for researchers, however, sifting through the available data is an increasingly difficult challenge (Bergman 2001).

Bottlenecks in the research processes are becoming evident because researchers are now faced with making use of, and processing, this growing amount of data (Hall et al. 2009; De Roure and Goble 2009).

1.3.1. The Data Deluge in Earth and Environmental Sciences

In the earth-sciences, environmental sensor networks are being deployed to gather data in real-time across widely distributed areas for applications such as environmental and seismic monitoring. Examples of grand scale continental and intercontinental initiatives in data acquisition include:

- The National Ecological Observatory Network (NEON), and
- The North-East Pacific Time-series Undersea Networked Experiments (NEPTUNE).

NEON is a national observatory, funded by the U.S. National Science Foundation (NSF), to discover and understand the impacts of climate change, land-use change, and invasive species, on the ecology of the U.S.A. The NEON research platform consists of distributed sensor networks and experiments, linked by cyberinfrastructure software tools, to record and archive ecological data for at least 30 years. This long-term data will be used to determine ecological responses of the biosphere to changes in land use and climate, and on feedbacks with the geosphere, hydrosphere, and atmosphere (NEON 2008).

In contrast, NEPTUNE is the world's first large-scale cabled ocean observatory. NEPTUNE operates under the Ocean Networks Canada (ONC) collaborative effort and is funded by the Canada Foundation for Innovation and British Columbia Knowledge Development Fund. NEPTUNE gathers live data from a rich constellation of instruments deployed in a broad spectrum of undersea environments and then transmits the data from the seafloor to an archival system. This system aims to provide access to an immense volume of data, both live and archived, throughout the planned 25 year life of the project (NEPTUNE 2009).

The focus of this thesis is on Marine datasets, in particular, those observed from remotely sensed ocean observation networks. In Australia, the Integrated Marine Observing System (IMOS) (GBROOS 2008), is developing new sensor technologies and processing methodologies throughout Australian oceans, including those of the Great Barrier Reef Marine Park (Kininmonth et al. 2004).

The Smart Environment Monitoring and Analysis Technologies (SEMAT) project (Johnstone et al. 2008) is largely driven by the need to create a low cost intelligent sensor network system for monitoring aquatic and coastal environments, and importantly the analysis of that data into information which can be used for management and planning. In the USA, NOAA's Integrated Coral Observing Network (ICON) is a coral reef monitoring program, which uses satellite and remotely sensed data to provide early warnings and long-term monitoring of domestic and international coral reefs (NOAA-ICON/CREWS 2008).

There are many other endeavours aiming to correlate the sensor data streams. Initiatives such as the Global Lake Ecological Observatory Network (GLEON) and Coral Reef Environmental Observatory Network (CREON) aim to bring the environmental sensor data together in their prospective repositories. GLEON's observatories consist of remote sensor instrumented platforms on lakes around the world that sense key limnological variables and then send the data, in near-real time, to web-accessible databases, where researchers and the public can obtain the data from a common web portal (GLEON 2009). In contrast, the aim of the CREON group is to provide global data, specific to scientists and marine managers, so that marine life can be managed with the best informed decisions of the day. CREON is dedicated to bridging science and ICT in productive eco-informatic platforms (CREON 2008).

Importantly, the integration of the growing amount of sensed data, with other forms and types of data, is of immense value in the endeavour to create new knowledge. Therefore, the prompt and usable availability of such real-time data, which can be integrated with data from other sources (e.g., satellite, models and/or historic data sets) to produce new information, is essential for both environmental managers and researchers. The development of a new approach to extrapolate information from diverse data sources is the focus of this thesis.

1.4. Data Acquisition and Integration Decisions in Coral Reef Studies

A complexity of environmental research is that raw sensor data is often collected independently and is often heterogeneous. Notably, similar information is often collected by different organisations but maintained in non-interoperable forms. For example, both NOAA and Australian Bureau of Meteorology (BOM)[2] gather data of environmental factors such as weather. However, the data is heterogeneous in terms of data standards, temporal/spatial resolution, etc.

[2] http://www.bom.gov.au

This heterogeneity works to impede data integration by individual researchers, and ultimately the discovery of new knowledge, which could come from merging the independent data sources.

Marine biologists and ecologists normally base their methodological and physical decisions about research strategies on four pragmatic categorisations of data collection (Olsson et al. 2008; Marshall and Schuttenberg 2006):

1) The sample size –
    a. What is the most effective sample size to use in a given study? For example, one sample per day at a set time for a small set number of days might be more optimal than one sample per minute for an extended period of days/months when determining how to measure solar intensity. The decision is typically dependent on the specific study.
    b. Capacity consideration - is all the data to be collected imperative or are there redundancies?
2) The quantity and types of data –
    a. The volume of data – how much data becomes too much in terms of numerical analysis and the ensuing processing requirements? The answer is dependent on the computing facilities and time involved in processing the data, as opposed to the correct sample size required for a comprehensive study.
    b. The ecological consideration – what is the environmental parameter measured that is conducive to the question being posed and are there proxies that can be used instead? For example, satellite sensed ocean chlorophyll is sensed by colour and is a proxy for the amount of nutrient in a location. Although it is not a fully accurate proxy, the chlorophyll level is currently the most cost effective and feasible wide scale, remotely sensed proxy for ocean nutrients in use (AIMS 2007a).
3) Physical management and referencing of data –
    a. What needs to be considered for the future use of the data? The decision of how much data to collect and for how long is not only based on its immediate application but also its future use.

b. How much metadata, and to what detail, is required when acquiring, archiving and storing the data? When extra information is added to the raw data (e.g., provenance metadata, access and control metadata, etc.), the capacity and volume increases considerably.

c. Should consideration be given to further studies or just collected for one specific study? Because a larger sample size (e.g., samples at hourly intervals instead of daily) may be needed in future studies, to gather the data in accordance with opportunity costs may be more effective. For example, more than one study might possibly benefit from the data, so if collection decisions were based on the highest common denominator, even if the initial costs are raised, it would ultimately be more cost effective. However, knowledge of future studies would be required to make expenditure decisions based on the balance of valuable data versus wasted resources (e.g., time, money, storage capacity, etc.).

4) Future compatibility of data value -

a. What data format will best fit the needs of the present and future research? Are the current formats sustainable of are they reliant on proprietary software?

b. In contrast to open source data, proprietary data, or closed source data, is not accessible to the wider scientific community. How does the owner of the data source recoup the costs of data collection, if published openly? Also, to maximise the value of the data, when is the most prudent time to permit open availability to the wider community (often a consideration of the government)?

c. Who is responsible for the maintenance of the data, once published in an open forum? Specifically, to ensure that policies of quality assurance are implemented and the integrity and quality of the data are upheld.

Currently researchers use these categories, or some variations of them, to design their research plans. However, these limitations in data collection scale are being challenged by remote sensor networks and remote sensing (data deluge). If better reuse of data, and access to data repositories was possible, it would be viable to extend the scope and extent of coral reef research. Data from individual studies could be synthesised with related data from other studies to build a

more complete analysis. This research aims to provide a basis for the reuse of data in an efficient manner by integrating and posing questions of the data.

1.5. Eco-informatics – Techniques in Cross-discipline Research

Many of the most influential papers in coral reef science of the past few years have been "synthesis" papers which aggregate long-term observations into new hypotheses and conclusions. One well known example of this new class of science would be the recent Inter-governmental Panel on Climate Change reports (IPCC) (2007). The IPCC was established to provide the decision-makers, researchers and the public with an objective source of information about climate change and the causal factors of climate change. The IPCC is an international, intergovernmental, scientific body and the reports it provides are stated to: "assess on a comprehensive, objective, open and transparent basis the latest scientific, technical and socio-economic literature, produced worldwide, relevant to the understanding of the risk of human-induced climate change, its observed and projected impacts and options for adaptation and mitigation." (IPCC 2007). The IPCC reports are the product of many scientists working together to review and synthesise research on climate change. Research that includes studies conducted in ecological domains such as coral reefs to find correlations in data through observational hypothetical.

Eco-informatics can be defined as the combination of multiple environmental datasets and modelling tools to test ecological hypotheses and derive information. There are a variety of technologies being developed to enable e-Research requirements, such as semantic technologies, Grid computing, and scientific workflows. When enlisted in an eco-informatics application, each of these technologies has particular characteristics and strengths to meet the needs of the modern e-Researcher. These needs include, but are not limited to, the ability to process large quantities of data from diverse origins and in differing formats, the simplification of data integration and analysis and scalable, flexible automation of the processes undertaken (Gil et al. 2007; Hall et al. 2009; Goble et al. 2006).

1.6. The Semantic Reef Project

The Semantic Reef project is an eco-informatics application that is focused on the development of an automated data processing, problem-solving and knowledge discovery system to better understand and manage reef ecosystems.

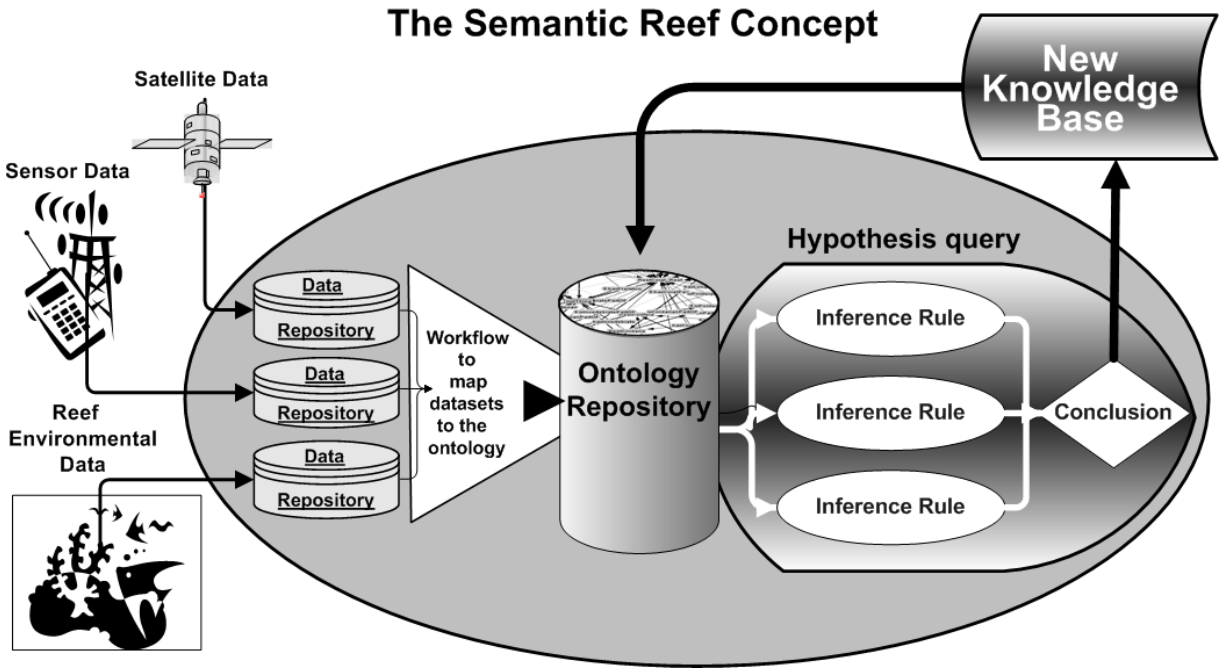


Figure 1.2 – The Semantic Reef workflow concept.

The goal here is to harness enabling technologies, specifically; Semantic Web and scientific workflows in an integrated hypothesis-based research tool (refer to Figure 1.2). Semantic Web technologies are inherently computer-centric with concentration in software, data and application layers and they connect currently disconnected data and integrate it in ways that can be manipulated by the computer. In contrast, workflow technologies, are both human and computer oriented, that is, they enable people to make the connections between the different technologies, software and hardware from diverse domains. As DeRoure (2004) concluded "software is the power behind the scientist, and the scientists are the power behind the software" which concisely sums up the symbiotic relationship the modern researcher has with the modern methodologies and tools they use. The Semantic Reef project is an example of just such a tool.

Implementations of semantic technologies within e-Research are prevalent in disciplines such as the Medical and Life Sciences, but are still emerging in the Earth and Ecological Sciences (Goble 2005). The growth in the research and development for the semantic technologies and tools has been driven predominantly by Genomics and Medicine. However, in Ecology and Marine Science these technologies have been only gradually adopted. Nevertheless, data collection methods are changing, remote sensing technologies are developing and deployed sensor networks on reefs are proliferating and growing, therefore, the need for better management of the data produced has arisen. The Semantic Reef project is a proof of concept that semantic technologies

can be used in non-typical ways to introduce new methodologies for research in the marine ecology.

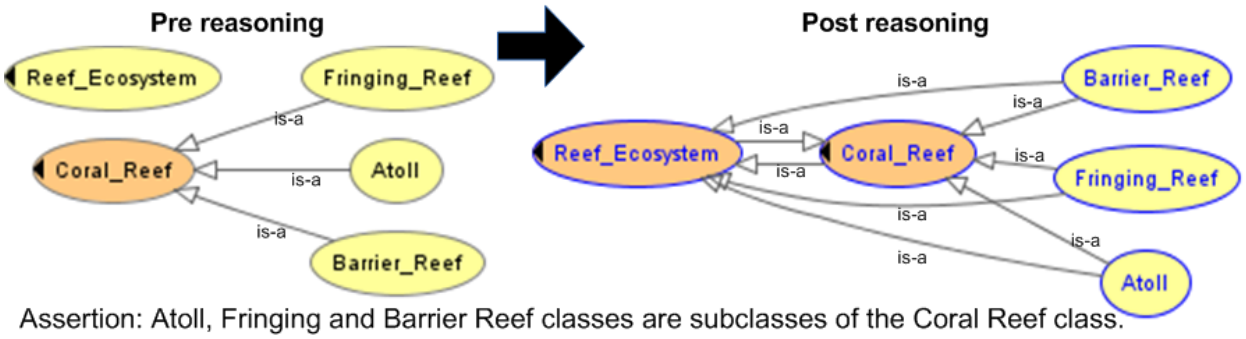
1.6.1. The Technologies

1.6.1.1. Semantic Web Technologies

The Semantic Web (SW) is an initiative of the World Wide Web Consortium (W3C) (W3C 2009c) aimed to create a "Web of linked data", as opposed to the current "Web of documents". Specifically, the SW is an evolving development in which the semantics of the information and services, currently available on the Web, are well defined, making the web content understandable to the computer and therefore automatically processable by the computer.

Data integration is a primary motivation in the development of Semantic Web technologies and is at the heart of the Semantic Reef project. Semantic Web technologies are an emerging area of computer science that support automated processing of information. Ontologies lay the foundation of technologies and are "specifications of a conceptualisation" (Gruber 1993) to give context and meaning to the data available to the computer (Antoniou and van Harmelen 2004). Ontologies describe "things" that exist within a domain, whether they are abstract or specific. A concept is modelled by specifying the domain's vocabulary and the terms that describe the entities within it and then the relationships that exist between entities are explicitly defined using axioms and restrictions. The interpretation of the concept, for use by the computer, is then constrained, which makes the concept machine-processable (Guarino 1997). Herein, a set of reusable ontologies have been developed to describe to a computer the concept of, and the relationships within, a coral reef ecosystem.

Ontologies can bridge disparate data held in data silos or make sense of the numerous pieces of information available on the Web (Antoniou and van Harmelen 2008). Currently, other approaches amalgamate heterogeneous data sources, including data-warehousing and data-mining. However, the application of ontologies for these integration tasks is potentially more effective as they can resolve the semantic conflicts of definitions, which invariably arise from diverse schematic sources (Wache et al. 2001). The Linked Data movement is a seminal example of applied semantic technologies to enable data integration between Web-based data resources (Berners-Lee 2007). The goal of the Linked Data movement is to make online data more semantically aware via annotation so access to the data can be automated.



Assertion: Atoll, Fringing and Barrier Reef classes are subclasses of the Coral Reef class.
Assertion: Coral Reef is equivalent to Reef Ecosystem.
Inference: Atoll, Fringing and Barrier Reefs are subclasses of a Reef Ecosystem class

Figure 1.3 – An example of automating equivalencies and subsumptions with DL

The level of granularity of an ontology can range from informal or "lightweight" to formal or "heavyweight" (Lassila and McGuinness 2001). Domain vocabularies, thesauri or taxonomies are examples of lightweight, informal ontologies. In contrast, the heavyweight variety applies formal logical definitions, which make possible the automation of classifications of concepts and the interference of new information (Gomez-Perez et al. 2004). Because the data and information is interpretable by a computer, through the informal or formal ontological definitions, the machine is enabled to make intelligent decisions. Decisions based on conclusions are derived through logic systems such as inference rules and Description Logics (DL) (Antoniou and van Harmelen 2004).

Inference rules written in syllogistic form offer a platform for hypothesis testing. The Semantic Web Rules Language (SWRL) is the proposed inference rule language for the Semantic Web (Horrocks et al. 2004). Through a combination of the Web Ontology Language (OWL) and the rule Markup Language, SWRL enables an environment that supports suppositions (Horrocks et al. 2004; O'Connor et al. 2005). SWRL manages inference using Horn logic, a subset of predicate logic, or First-Order Logic (FOL), where syllogisms are presented and result in automatically inferred conclusions.

DLs are sets of logical statements to describe relationships between entities of a concept. That is, the axioms and restrictions of an ontology which constrain the parameters of that concept can be used by a reasoning engine to infer equivalencies, assumptions and subsumptions. Figure 1.3 depicts a simplified example of how equivalent relations and subsumptions can be automatically classified. Firstly, on classifying the ontology with the reasoner, an instance from one database (i.e., "reef ecosystem") is automatically classifed as equivalent to an instance of another (i.e., "coral reef"). Further, subsumption is illustrated with the "coral reef" sub-classes: atoll, fringing and barrier reef classes (Figure 1.3), which are automatically subsumed to belong to

the "reef ecosystem" class, simultaneously, based on the given axioms and restrictions (i.e., assertions).

A simple food web is a more complex example, where the axioms and properties that define the class are used to classify the ontology. For instance, the property "eats" can be asserted to classes with restrictions on what kind of food (e.g., a carnivore "eats" ONLY meat). On reasoning over the ontology a member of an omnivore class would be automatically subsumed to also belong simultaneously to the herbivore and carnivore classes. In these cases, the computer automatically connects the latent or "hidden dots" in a semantic Knowledge Base (KB) by classifying the classes and instances of the ontology.

1.6.1.2. Scientific Work Flows

Scientific workflow technologies and tools are adaptive software programs to capture complex analyses in a flow of which the data is taken through one analytical step after another (Altintas et al. 2004). The Kepler system is an open-source scientific workflow tool and is the software chosen for the data flow implementation of the Semantic Reef architecture. Each workflow step is represented by an "actor" in the Kepler system and they provide access to the continually expanding amount of geographically distributed data repositories, Grid computing resources, and workflow libraries (Ludäscher et al. 2006).

Grid technologies permit decentralised management of resources that can be simultaneously accessed and attained from geographically separate organisations (Foster et al. 2001). The Grid middleware provides Web and Grid services access to the data resources, handle security, data movement controls and resource monitoring and discovery services as required. Data grid tools are available via scientific workflows and offer retrieval services to data contained in the repositories of research partners (Foster 2002).

Kepler workflows are employed to automatically process raw data and/or web available data via a series of workflow steps, and pass the results to the Semantic Reef KB. Specifically, the workflows collect both raw data from remote sensors as well as existing data from archives and repositories by enlisting both Web and Grid services and ultimately, the KB could be filled with relevant data available from diverse sources. For example, physical parameters such as ocean temperature, salinity, nitrogen, pH and bathymetry information, as well as biological data such as coral, algae and fish stocks. Knowledge may then be derived from the data by questioning semantic correlation and analysis using DL and inference rules. Hypothesis questions or alerts can

then be posed, such as, finding the tipping point between a healthy reef and a dying reef, or alerting to events such as coral spawning or bleaching (refer to Figure 1.2).

1.7. Semantic Reef Project - Aims and Objectives

1.7.1. The Research Aims

The Semantic Reef project is a knowledge representation platform to allow researchers to combine disjoint data into a single KB, to pose questions of the data. Scientific workflows are used to access and retrieve remote sensor data, data available via the WWW and/or data available on data-grids, then combined with data already integrated into the KB. When coupled to the datasets, the reef ontologies developed herein, derive inferences from data to "ask" the system questions for semantic correlation, synthesis and analysis.

The hypothesis of this dissertation can be simply enunciated as:

To assess the feasibility of using semantic inference in a hypothesis tool to facilitate research on coral reefs by inferring information and/or knowledge from multi-scale, distributed data.

1.7.2. Research Objectives

The objectives for this work are as follows:

- To investigate the capabilities and synergies of semantic technologies and scientific workflows as methods for data integration.
- To investigate new means in hypothesis modelling and design to enable marine researchers to make efficient use of the data from new collection efforts such as remotely sensed networks. The new means should allow a new research potential to resolve or answer questions such as the effects of climate change on coral reefs.
- To develop an ontology framework that can be reused for any coral reef and is independent of the line of query, the location and/or the data.
- To bridge and combine complex collective knowledge, which is currently held in various data forms within separate research institutions, into one KB for use in hypotheses-driven research.

- To successfully integrate the emerging Semantic Web technologies with scientific workflows into an architecture which allows marine researchers to flexibly pose observational hypotheses based on a richer source of data and information.

1.7.3. Research Contribution

The Semantic Reef model is a research case study that combines scientific workflows, Semantic Web technologies, FOL and propositional logic, in a KR system. The particular combination used within the project's architecture is an exemplar of future ways to manage rich data sources in a more productive manner. The project offers the following contributions:

- This thesis is the first known example of the application of Semantic Web technologies and workflows combined to integrate data with the purpose of posing observational hypotheses or inferring alerts in the coral reef domain. In fact, it is one of the very few examples of eco-informatics of this type known.
- A significantly broader scope of available information about the coral reef domain is made available to observational hypotheses through the integration of scientific workflows technologies in the model.
- The development of a new model to the proof of concept stage, which can be assessed as a tool to analyse disparate datasets and to discover new knowledge by adopting hypothetically driven research processes.
- Successful testing of the model (the Semantic Reef) as a proof of concept to determine its viability as a tool to assist coral reef biologists in the prediction of events such as coral bleaching or observational hypotheses to discover phenomena in the data.

1.7.4. Research Constraints and Assumptions

The restrictions found during the course of the study are as follows:

- Data availability – Gaps in the available data currently exist, which hinder hypotheses. To infer and conclude over data instances the relevant data must be available for import to the KB. If the data does not exist (i.e., has not been or is not being collected), or the data is from a closed source, then it cannot be used in the hypothesis.
- Implementation constraint – the Semantic Reef architecture is a proof of concept that has implemented emerging technologies as they are being developed. A web portal is a

component of future work to allow scientists to create the rules (hypotheses) and specify the required data.

- Domain expertise – Hypothesis design when crossing disciplines is a complex undertaking. Collaboration and open communications were required to bridge disciplinary knowledge diversity, particularly when translating hypotheses from specifications of domain experts to propositional inference rules. Well managed communications are integral to the creation of any future hypotheses.

1.7.5. Research Approach and Chapter Synopsis

Chapter 2 presents a detailed literature review on the e-Research paradigm. The problems being faced by researchers and the techniques used to overcome challenges in modern research paradigms are discussed. Specifically, how and why the data deluge is an emerging problem, and what current technologies and standards are evolving to overcome it. Additionally, an overview of previous and current work undertaken in the areas of Semantic Web and workflow development is presented.

Chapter 3 describes the design approach taken in the ontology development of the Semantic Reef architecture. The hierarchy of ontologies within the KB consist of reusable domain ontologies and usable application ontologies. They range in complexity from informal taxonomies through to formal ontologies. The informal taxonomies describe reef system concepts such as community stock and environmental characteristics (e.g., percentage of hermatypic corals, temperature, salinity level, etc.). In contrast, the formal ontologies describe the ecology and the interrelationships between elements of a coral reef. For instance, the tolerance and interdependence of reef organisms, like corals, to physical parameters like temperature. The informal lightweight ontologies are imported into the more complex formal ontologies to create a "ground-up" importation architecture, and with each level of granularity the designated purpose of the ontology narrows.

Chapter 4 details the validation of the KB. The validation process involved a reverse-hypothesis approach to ground-truth the system against historic events. The reef ecosystem ontologies were coupled with historical datasets from the 1998 and 2002 mass coral bleaching events on the Great Barrier Reef (GBR). Then, inference rules were developed to model current analysis metrics used in the prediction of coral bleaching events. The results from running the rules showed the inferred conclusions were the same as the historical outcome, thus validating the accuracy of the KR system as a prediction tool.

Chapter 5 provides demonstrations that show the differences semantic technologies, combined with scientific workflows, offers research via observational hypothesis. The demonstrations begin by substituting historic Sea Surface Temperature (SST) data with near real-time SST data streamed, via Web service "actors" in Kepler, from the Australian Institute of Marine Science (AIMS) data centre (GBROOS 2008). Then, data from a variety of sources is integrated to the system for arbitrary hypotheses. Finally, a demonstration in the automated classification capabilities of the semantic inference is provided by automatically classifying reefs to belong to a specific reef-type, such as, by location or by community composition.

Chapter 6 presents a detailed performance analysis to explore the functional and practical application of the Semantic Reef architecture as a hypothesis and/or predictive tool. The analysis is simulated in a standard desktop environment, which is indicative of a researcher's computing paradigm while designing hypotheses.

Chapter 7 concludes the thesis by summarising the research contribution, followed by a description of future work and directions.

Chapter Two

Review of Literature and Methods

2.1. Introduction and Chapter Synopsis

The Semantic Reef project is an eco-informatic e-Research application. E-Research refers to the amalgamation of research techniques, data and people with Information Communication Technologies (ICT), to enhance technical capabilities or solve problems in the acquisition of knowledge. Modern research practices are advancing as different enabling e-Research technologies are combined to form synergies that can improve previous research methodologies and/or enable new research processes (Taylor et al. 2008). The development of new e-Research technologies and solutions is driven by the requirements of the researchers have Virtual Research Environments (VRE), specialised hardware and data integration methods, more efficient data collection, storage, analysis and synthesis.

However, as modern research processes and practices advance, new problems emerge. One such problem is the exponential growth in scientific data caused by advancements in scientific instrumentation and data gathering techniques. To better manage the onslaught of scientific data, technologies are in development, which among others include the Semantic Web, Grid computing and scientific workflows. The Semantic Reef architecture combines components of these technologies and is designed to explore different methods that can potentially alleviate the bottlenecks in the data analysis and synthesis processes arising from the massive growth in data.

In this chapter, a discussion of e-Research is presented, including its evolution and its importance to the modern research environment. As well, technologies that offer solutions to problems modern researchers are facing are detailed. Following this, current e-Research techniques, specific to the marine science environment, are explored, including related environmental projects. To conclude, the Semantic Reef project, the topic of this thesis, and the solutions it potentially offers to coral reef ecology, is introduced.

2.1.1. E-Research - The Definition and Evolution

The e-Research concept evolved from the e-Science movement. The term e-Science was coined in 1999 by Professor Sir John Taylor (former Director-General of UK Research Councils)

who claimed: "e-Science is about global collaborations in key areas of science and the next generation of infrastructure that will enable it" (NeSC 2009; Hey and Trefethen 2002). The term initially described a large funding initiative that began in November 2000, but has since evolved to depict any scientific project that is computationally intensive and implemented in highly distributed networking environments. Alternately, in the USA, the term cyberinfrastructure is typically used to define e-Science projects. Both terms are now commonly used interchangeably and refer predominantly to the High Performance Computing (HPC) and Grid computing infrastructure that make possible the analysis and visualisation of massive sets of data (De Roure and Hendler 2004; Hey and Trefethen 2003b). E-Research expanded the e-Science concept to bridge scientific disciplines and envelop all forms of exploration, research and development in a more holistic framework (Hey and Trefethen 2003b; Taylor et al. 2008). All three terms have now become synonymous in meaning and all refer to the enablement or enhancement of capabilities in knowledge acquisition by creating synergies through the combination of people and machines (Goble et al. 2006; De Roure et al. 2004).

For the purpose of this thesis, e-Research expands on the e-Science concept by including technologies that support the common processes of all research domains, as articulated in the UK Joint Information Systems Committee (JISC) definition:

> *"e-Research refers to the development of, and the support for, information and computing technologies to facilitate all phases of research processes. The term e-Research originates from the term e-Science but expands its remit to all research domains not just the sciences. It is concerned with technologies that support all the processes involved in research including (but not limited to) creating and sustaining research collaborations and discovering, analysing, processing, publishing, storing and sharing research data and information. Typical technologies in this domain include: Virtual Research Environments, Grid computing, visualisation services, and text and data mining services." (Allan et al. 2004)*

New generations of scientific instruments are capable of producing or gathering massive quantities of data. In fact, as discussed by Hey and Trefethen (2003a), instruments such as synchrotrons, particle accelerators, diffractometers, distributed sensors, satellites, etc., are producing data into the Petabytes[3] per year. Consequently, the emerging requirements for the modern researcher include the ability to share, manage, control, access and analyse the data,

[3] A Petabyte (PB) is a unit of information or computer storage that equates to approximately one quadrillion bytes, or 1024 terabytes.

information, new ideas or new knowledge, which these data acquisition activities produce (Hey and Trefethen 2003a; De Roure et al. 2004; Hall et al. 2009).

2.2. Modern research requirements

The e-Research community is working to develop new methods that offer solutions to the changing requirements of the modern researcher and/or scientist. These needs include, but are not limited to:

- The capability to process large quantities of data from diverse origins and format;
- The facilities to share both tangible and intangible resources;
- The facilities to maintain collaborative, dynamic environments;
- The ability to simplify data integration and analysis; and
- The provision of scalable, flexible automation of the processes (Gil et al. 2007; Hall et al. 2009; Goble et al. 2006).

Three major technological fields are currently in development to fulfil some of these five areas of modern researcher's needs: VREs, which employ both collaborative and workflow tools; hardware requirements; and data integration methods.

2.2.1. Virtual Research Environments

VRE's are the collection of technologies that enable or improve the working environment for the modern researcher. A VRE is required to support simplicity while simultaneously streamlining all stages of the research methodologies and processes, from the initial documentation of the theory or hypothesis to the physical undertaking and through to the final publication of results (Gil et al. 2007). Further, the VRE is also required to be independent of the research method adopted, whether it is statistically based, *in silico, in vitro, in situ* or other type.

A VRE facilitates the sharing of digital and electronic resources in a structured, orderly and secure fashion. These resources include data, digital files, visualisations, publications, computational power and storage, among others. The VRE is a collaborative working and research environment that combines all aspects of the workflows in a transparent manner. Because the VRE allows such flexible and diverse collaboration it is fully suited to fostering multi-disciplinary and interdisciplinary work, and is by no means limited to the speciality of the one researcher (Hendler 2003). Also, collaborative equity facilities are another important component of a VRE because they allow the individual contribution of researchers to be acknowledged equitably, and therefore

can avoid potential problems of intellectual property and credit assignments at both organisational and individual researcher levels (Waldrop 2008).

The VRE fosters the collaborative nature of science. Scientists communicate findings in an open environment so that others may learn, disseminate or scrutinise the results, which can ultimately expand current knowledge. Goble (2006) recently summarised the collaborative concept succinctly:

> *"Scientific progress increasingly depends on pooling know-how and results; making connections between ideas, people, and data; and finding and reusing knowledge and resources generated by others in perhaps unintended ways. It is about harvesting and harnessing the "collective intelligence" of the scientific community."*

The ARCHER project[4] in Australia and the myExperiment project[5] in the UK are examples of VREs. These projects apply current and emerging technologies to promote communication and academic discussion and sharing, and ultimate reuse, of data and information. The tools, technologies and resources available to researchers, via these facilities, include support, access and management of large datasets, enriched with metadata, from distributed repositories. The VREs offer collaborative functionality for sharing documents, publications and scientific workflows and also the new information derived from the data analysis processes (Atkinson et al. 2008; De Roure et al. 2009).

2.2.2. Hardware Requirements

To deploy e-Research experiments and to support endeavours, such as the ARCHER and myExperiment VRE's, a sound, reliable, extensible hardware infrastructure must be in place. More specifically, the researcher's hardware requirements must include a high bandwidth infrastructure for data throughput, High Performance Computing (HPC) and the support and tools to automate integration, management, visualisation and analysis of the data (De Roure and Goble 2009).

Internetworking technologies lay the fundamental physical support for e-Research. Modern research and development activities have instigated the forward movement of networking technologies and are focused predominantly on scalability, fault tolerance, latency and the growth in available bandwidth. More explicitly, the higher the bandwidth the greater the actual throughput and the greater the throughput the more support for data transfer, virtual working environments and collaborations on a global scale (Hey and Trefethen 2002; Hey and Trefethen 2005).

[4] http://archer.edu.au/
[5] http://www.myexperiment.org/

The Grid computing paradigm is regarded as e-Research's infrastructure. Many scientists are drawn to Grid computing for HPC resources, to support data management and analysis across sites and organisations. Grids provide support for distributed computation by managing the execution of complex job workflows and facilities for robust efficient file management, transfer and sharing (De Roure et al. 2004). Grid computing is discussed in greater detail in section 2.4.

2.2.3. Data Integration Requirements

The means to physically share, move and maintain data more efficiently does not resolve all the modern researcher's requirements and so new data integration methods are imperative. The endeavour to create collaborative research environments requires a foundation of Web and Grid services and global high-speed research networks (Hey and Trefethen 2002; Goble et al. 2006). However, data and information management is also an integral function because interoperability is essential when data from different platforms and formats converge. As predicted in Moore's Law[6], the hardware and networking resources that make up the underlying infrastructure are in perpetual growth, and improvement phases and technological developments in data integration and management are not keeping up as the data continues to increase (Gil et al. 2007; Hall et al. 2009; Hey and Trefethen 2003a).

2.3. The Data Deluge Problem

2.3.1. Data Gathering Instruments

Researchers are faced with a growing amount of data to process. The imminent influx of new data and information, appropriately dubbed the "data deluge" is changing data management and processing requirements (Hey and Trefethen 2003a). This increasing flood of data is growing exponentially with the large number of deployed, or soon to be implemented, data collection instruments such as accelerators, sensors and satellites, and is again escalated when the outcomes of experiments and simulations also contribute to the deluge. Consequently, bottlenecks in the data processing and analysis phase arise as the volume of raw data and Web available data increases. This bottlenecking is compounded because current data processing methods still involve manual manipulation and human intervention, and manual data processing is restricted by limitations on capacity and time. Therefore, it is becoming progressively more difficult for the researcher to effectively keep up as the quantity of data increases (De Roure and Goble 2009; Hall et al. 2009).

[6] The exponential increase in computing power, networking and solid-state memory

There are two categories of data collection instruments in the production of data: those that scale-up and those that scale-out (Szalay and Gray 2006). Instruments that scale-up refer to the machines deployed that, once turned on, produce massive amounts of data, but the quantities of the data are always approximately the same. These machines are usually the product of large scale projects, possibly from the efforts of multiple countries and involving thousands of people. For example, the Large Hadrons Collider (LHC), which is predicted to produce approximately two Petabytes of data per year, is the product of a global collaborative effort involving 80 different countries and over 8,000 scientists (CERN 2008). However, because the massive quantity of data the LHC produces will remain reasonably constant, it is classed in the category of a scaling-up data production instrument.

In contrast to the scaling-up instruments, which produce a sizeable yet finite amount of data, the instruments that scale-out produce less data but can be deployed at an exponential rate. The scaling-out instruments are deployed singularly because they are usually "stand-alone" machines that are tasked with a set application, method or objective (e.g., a sensor that measures temperature). Deployment costs are much lower with these types of instruments because individually they are cheaper than the scaling-up variety. The quantity of instruments included in a data gathering project is usually driven by the immediate budget of a specific project and also, more can be added at any time as the research requirements warrant or as budgets allow. Consequently, because of the low costs and flexible nature often allowed by the scaling-out data production, it is possible for an exponential growth in deployed instruments, which in turn results in an exponential growth in the data they produce (Szalay and Gray 2006).

2.3.2. Data on the World Wide Web

The World Wide Web (WWW or Web) is the major source of the collective knowledge of humanity and is continually growing as more documents, data and information are added (Berners-Lee, Hall, Hendler, Shadbolt et al. 2006; Hendler et al. 2008). The WWW was invented in 1989 by Tim Berners-Lee and the initial proposal sought to provide a distributed hypertext environment, which had the potential for scientists to share and distribute information more easily. The original design included the hypertext transfer protocol (HTTP), the hypertext mark-up language (HTML) and the first Web browser, which were the protocols and software developed to more easily share documents (Berners-Lee 2000b). The initial proposal offered benefits such as: accessibility from any enabled network connectivity; ease of use; ability to bridge across platforms; and an open source release agenda (Berners-Lee 1999; Hall et al. 2009). The open source nature of the Web,

including its original technologies and the many ensuing protocols and technologies, was a major reason for its phenomenal success and colossal growth (Hall et al. 2009). The Web is the integral component in the paradigm shift of how the world now researches, lives, works and communicates (Berners-Lee, Hall, Hendler, Shadbolt et al. 2006; Hendler et al. 2008).

The number of Web pages today still show amazing growth rates (Adamic and Huberman 2002). As at March 2009, the indexed Web contained more than 23.94 billion (2.4×10^{10}) pages, according to the "World Wide Web Size" calculator[7], which is a project from the Netherland's Tilburg University. The calculator applies algorithms to quantify the size of the Web by estimating the numbers of pages indexed by major search engines, such as Google and Yahoo Search (de Kunder 2006). Conversely, an announcement in July 2008 on the official Google Blog[8] declared: "systems that process links on the Web to find new content hit a milestone: one trillion (the order of 10^{12}) *unique* URLs on the Web at once". Further, according to the US Census Bureau's International population clock[9], the world population as at May 2009 was 6,768,167,712 (6.7×10^{10}). So comparatively, there are approximately 150 web pages for every person on the planet, which indicates the daunting magnitude of the Web. Consequently, the exponential growth of the Web is an example of a data source that is scaling-out and contributing to the problematic data deluge (Huberman and Adamic 1999; Szalay and Gray 2006).

Importantly, information and data available via Web resources is commonly not appropriate to employ in research, because of concerns about quality assurance and data integration. The open nature of the Web is one reason for its popularity (Hall et al. 2009). Because the Web is a forum where anyone is (technically) allowed to say anything about anything (W3C 2009a) full quality control and assurance of Web available content is impossible. The documents and information available for researchers to sift through and utilise in their studies, given the rapid growth rate of the Web, is beyond challenging; it is overwhelming. Researchers are faced with managing the validity of external sources, because there is no governing quality restrictions on what is publicly exposed on the Web and therefore the integrity of the content available is mostly questionable (W3C 2009a; Allemang and Hendler 2008).

In addition, with approximately one trillion pages containing available, and potentially useful, information and data, much of it is lost due to the disparate nature of the resources. Data is held in repositories, data silos and backend data-bases, which collectively was dubbed the "deep Web" by Bergman (Bergman 2001) in a discussion about the true depth of the WWW. In most

[7] http://www.worldwidewebsize.com/
[8] http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html
[9] http://www.census.gov/ipc/www/popclockworld.html

cases communication between these data sources is not possible, nor can the data be easily merged, due to disparate formats, language barriers and heterogeneous platforms. Notably, the communication between data sources is necessary for research conducted to search for phenomena and correlations in Web available resources. At best, the researcher manually "screen scraps" what data they require and systematically converts or transforms the data for integration into a study. A possible solution to the problem of disjoint and disparate information and data resources on the Web is Semantic Web technologies and the linked data movement.

2.4. E-Research Enabling Technologies

The technologies most typically used in e-Research applications each have key functions that create synergies for better solutions to the requirements of the modern researcher. Typical technologies include Semantic Web technologies, Grid computing, scientific workflows, text and data mining and integration services, VRE enabling tools and visualisation services. The main technologies applicable within the scope of this thesis are Semantic Web technologies, Grid computing, and scientific workflows.

2.4.1. Semantic Web

The Semantic Web is an initiative of the Worldwide Web Consortium (W3C) (W3C 2009c). From the inception of the WWW, Berners-Lee (1990) described it as a Web of documents that will evolve to a Web of data – a Semantic Web (Berners-Lee 2002). Berners-Lee's vision of the Semantic Web is defined as an extension of the current Web in that it provides a framework to autonomously share and reuse the data available on the Web via software agents (W3C 2009c; Berners-Lee et al. 2001). The technologies make available contextual information about the data, and thus make the data machine-understandable and ultimately machine-processable. (Antoniou and van Harmelen 2008). This form of machine processing enables the automation of tasks such as data fusion and data integration. Hence, the Semantic Web technologies offer great potential in the federation and/or linking of data stored on the Web and the "deep Web" (Bergman 2001).

The Semantic Web links data so it can be accessed, reused and/or manipulated more readily by the machine. The concept creates links between data, rather than simply inputting data on the Web. When the machine has information about a specified concept, it can explore the Web to find other related concepts and this form of linking is analogous to a neural network (Berners-Lee 2007). Currently, the Web provides links between pages (hypermedia) that are predominantly designed for human consumption. The Semantic Web augments this with pages designed to

contain machine-readable descriptions of the specific content and resources of the Web pages from a website. These documents can then be linked together to provide information to the computer which show how the terms in one website relate to those in another (Hendler 2003). For example, if additional contextual information is available, such as defined equivalency statements, the computer may "understand" equivalencies and synonymous concepts. For instance, a keyword search on information about Staghorn coral would retrieve only pages containing "staghorn" and "coral" using the current methods of searching Web content. Notably, both Staghorn coral and *Acropora*, which are the common name and the scientific name of the coral species, respectively, are terms commonly used interchangeably. If there is no contextual information to link these terms through their synonymous relationship, the computer would not be able to infer a connection. Hence, all pages that contain information on Staghorn coral but were referred to by its scientific name "*Acropora* would be missed in a standard keyword (Myers et al. 2008).

A major paradigm shift in how research is conducted will be possible through these technologies. Because the computer can contextualise data and information on the Web automatically knowledge can be extracted autonomously (Shadbolt et al. 2006). By adding well defined relations (e.g., synonymies, antonymies, homonymies, etc.) to the content available via Web computers can infer obvious or latent connections automatically, which will ultimately result in more intelligent search facilities.

2.4.1.1. The Semantic Web Architecture

The vision of the Semantic Web architecture as technological layers was presented by Tim Berners-Lee's in the "Semantic Web stack" or "layer cake" diagram (Figure 2.1), from a keynote given in 2000 (Berners-Lee 2000a; Antoniou and van Harmelen 2008). The languages and protocols, of the Semantic Web technologies have proceeded in steps. The basic building blocks of the Semantic Web architecture are the metadata, the languages, the ontologies and the logic and inference mechanisms (Antoniou and van Harmelen 2004; Fensel et al. 2002; Noy 2004; Allemang and Hendler 2008). Each of these steps builds on the previous layer and thus extend and exploit the features and capabilities of the layers below (Fensel et al. 2002; Berners-Lee, Hall, Hendler, O'Hara et al. 2006).

The foundation of the "stack" is the Unicode and Unified Resource Identifier (URI) protocols for the exchange of symbols and the standards to reference specific entities. Unicode is a universal standard encoding system for digital representation of human languages, symbols and scripts which allow computers to represent text in different writing systems (Unicode 1991-2009). In contrast, URIs provide a unique and unambiguous basis to locate resources by declaring compact

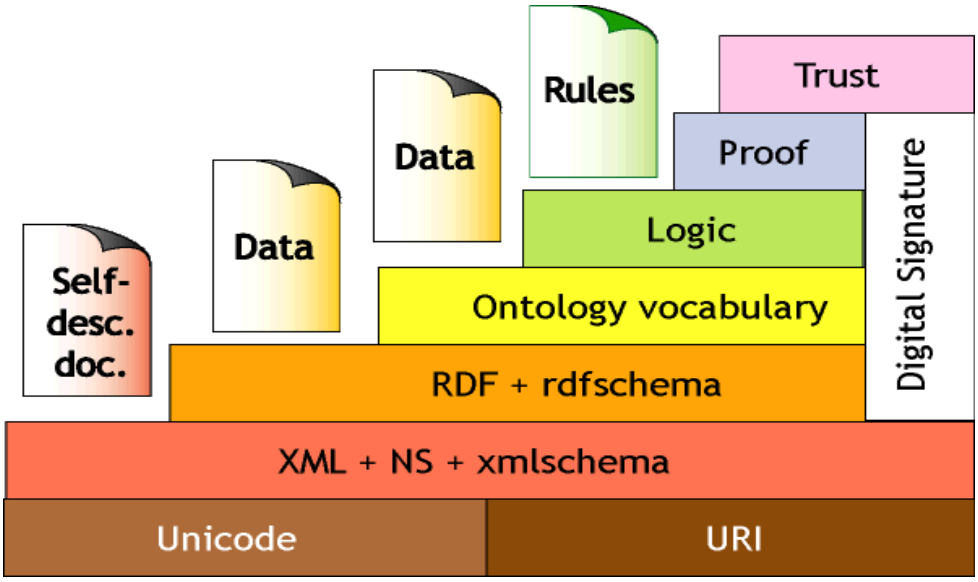


Figure 2.1 – The Semantic Web Architecture (Berners-Lee 2000a).

strings of characters that explicitly identify that resource (Berners-Lee et al. 2005). The URI specification, as part of the Web architecture, states that the URI does not need to be connected via a network to be declared. Rather, the resource can be abstract or physical (e.g., a phone number or home address). In fact, the resource can be anything that has an identity, so an unambiguously exact name (identifier) can be prescribed to represent it in a standard, uniform fashion. Two common types of URIs are the Uniform Resource Locator (URL), which is the absolute address (location) to a document or resource found on a networked infrastructure, and the Uniform Resource Name (URN), which refers to the name of the resource (W3C 2004a).

The eXtensible Markup Language (XML) and the XML Schema and namespace are at the next layer in the Semantic Web stack. XML is a set of clear rules for syntax for encoding documents electronically. The language allows the unique markup of elements to identify document structure by providing features for representing and interchanging information. Because the XML standard emphasises simplicity, generality and usability in describing information, it is the de facto standard for structured data on the Web. The XML Schema builds on XML as a method to compose XML vocabularies (Bray et al. 2008).

Although XML enables machine-processable information, it is insufficient to support semantic requirements alone. XML is a powerful surface syntax for structured documents but the descriptions are ambiguous to the machine because XML defines syntax and not the semantics of the descriptions. Explicitly, XML descriptions impose no semantic constraints on the meaning of this data and there are too many ways to describe the same thing. Therefore, since there is no real

meaning to support the structured data, the machine cannot make decisions, infer or imply anything autonomously with just straight XML. Alternately, the Resource Description Framework (RDF) and RDF Schema (RDFS) layer of the Semantic Web "stack" is the first step towards real semantic representation (W3C 2004b; O'Hara and Hall 2009).

2.4.1.2. The Semantic Layers

RDF is essentially a basic data model which uses simple descriptive statements called triples to represent resources (Manola and Miller 2004). A triple consists of a subject-predicate-object structure, which can be used to describe any resource, from people and places to web pages or Web services. All three components of the RDF triple are identified by an individual URI reference and in this format represent the metadata that describes a resource so it is consumable by the machine (Klyne et al. 2004). In fact, a triple can simply be described as three URIs, where the subject denotes the resource, and the predicate denotes the traits or properties of the resource and expresses a relationship between the subject and the object and is analogous to the grammatical rule of subject-verb-object in a sentence.

Data can then be stored, accessed and queried from flat files as triples, or structured by creating RDF Schemas (Brickley et al. 2004). An RDFS is a vocabulary description language that extends the RDF triple by adding additional semantic information such as basic class and property constructs to describe a concept (Lacy 2005).

The ontology layer is next in the Semantic Web hierarchy and provides functionality for richer definitions of concepts (Figure 2.1). RDF is the framework that can define vocabularies as objects by giving them properties and classes, whereas the Web Ontology Language (OWL) refines the vocabularies to a much greater extent through the more complex descriptive constructs available (McGuinness and Harmelen 2004). Constructs such as unambiguous cardinality constraints, property restrictions, existential and universal quantification and class and disjoint axioms (truisms) can be applied to describe concepts in more detail to the machine. Ontologies convey descriptions of worldly "things" stated in a fashion that is automatically computer-understandable and processable (Lacy 2005).

The ontology languages provide a means for the electronic creation of the machine processable descriptions. Today, the standard is OWL, which grew out of the earlier US DARPA Agent Markup Language plus the European Ontology Interchange Language (DAML + OIL) incipient standards (Horrocks 2002; van Harmelen et al. 2001). OWL Lite, OWL DL and OWL Full are the three current levels, or sub-languages, in OWL: and each level provides a different

degree of logical expressiveness (McGuinness and Harmelen 2004). Further, OWL 2 is a new W3C draft version of OWL, which extends OWL with new features such as extended data-types support, simple meta-modelling, extended annotations and additional property and cardinality constructs. As the ontologies written for this thesis are in the OWL standard, both RDF and OWL are described in greater depth in the following sections.

The semantic layers have moved forward from the original vision (Figure 2.1) due to implementation efforts prompted by both scientific and commercial communities (Hendler 2007). The evolution over the past decade has eventuated in subtle changes to the architectural vision. A more recent diagram of the layers in the Semantic Web "layer cake" (Figure 2.2) shows how the concentration has moved to the higher layers where the development of applicable technologies have been given more attention (Berners-Lee, Hall, Hendler, O'Hara et al. 2006; Hendler 2007; W3C 2009c). One such technology shown in the latest stack is the SPARQL Protocol and RDF Query Language (SPARQL) (pronounced "sparkle"), which is the new W3C standard for querying distributed triplestores[10]. (Prud'hommeaux and Seaborne 2008).

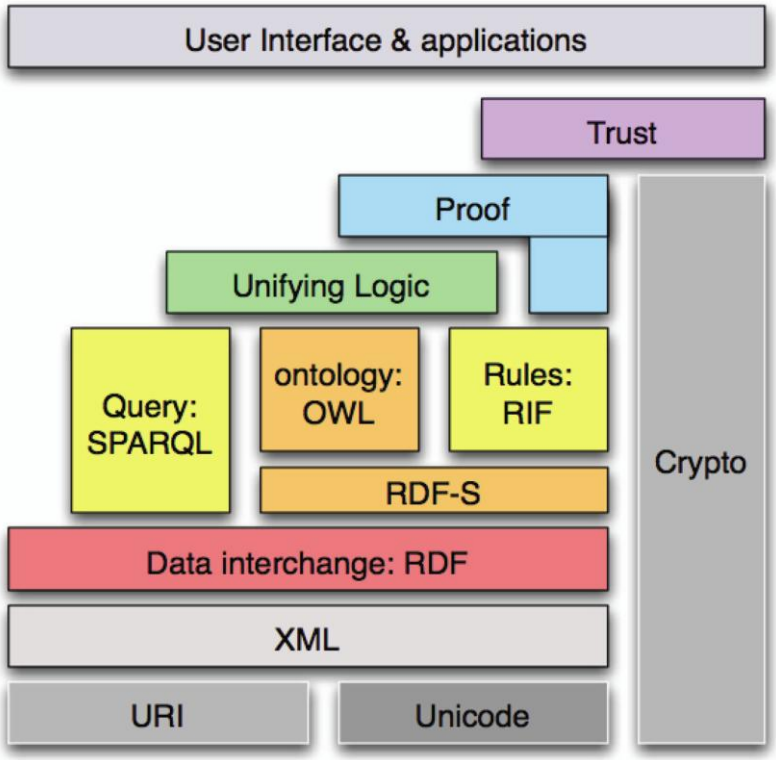


Figure 2.2 – The Semantic Web Architecture revised (W3C 2007).

[10] A triplestore is a purpose built database for the storage and retrieval of Resource Description Framework (RDF) metadata.

The logic and rules layers are located above the ontology layer and depict the languages and rules systems that enhance the lower layers. The logic layer in the earlier diagram (Figure 2.1) was purposely indistinct. Over time, as the development of the Semantic Web matured, the logic/rules layers became more definitive and required finer clarification; hence, the evolution of the semantic layers shown in Figure 2.2 (Berners-Lee, Hall, Hendler, O'Hara et al. 2006; Horrocks et al. 2005). One development is the Rule Interchange Format (RIF) for representing rules on the Web and linking rule-based systems together (Boley and Kifer 2008) and alongside OWL, the RIF is another extension of RDFS (Berners-Lee, Hall, Hendler, O'Hara et al. 2006).

The logic and rules layers offer more extensive rules and inference capabilities which build on the logic implementation of the OWL layers. The OWL ontology layer supports limited inference, such as, subsumption and classification with Description Logics (DL), a subset of First Order Logic (FOL), whereas the logic and rules layers support a wider variety of rules and inference systems. The goals of the RIF are to assimilate other rule-based formats such as Horn-clause logics, propositional logics and other higher order logics and production systems (Boley and Kifer 2008). Notably, many other formalisms are currently being adapted from the Artificial Intelligence (AI) field for the Semantic Web, including temporal (time-based) logic, causal logic, Bayesian reasoning, and other backward chaining probabilistic logics (Shadbolt et al. 2006).

The proof and trust layers (Figure 2.1 and Figure 2.2) support provenance and confidence in the results (Goble and De Roure 2002). The proof layer validates methods to verify the results of the rules and logic systems as well as the deductive process itself. At the trust layer, confirmation is given via digital signature of correct metadata, representation of proofs, proof validation and other recommendations of trusted agents on the level of security and privacy (Antoniou and van Harmelen 2008). Further, the revised vision of the Semantic Web stack (Figure 2.2) shows a new higher layer, which encompasses the entire stack and recognises the need for effective user interfaces and applications (Antoniou and van Harmelen 2008).

2.4.1.3. The Ontology

Ontologies are the foundation of the Semantic Web technologies and are defined by the *Oxford English Dictionary* as "the science or study of being". The term ontology originates from the field of philosophy which studies the nature of existence and essence (Gomez-Perez et al. 2004). Over the last decade, computer science has provided many definitions for ontology. One that appears most frequently in the literature is Gruber's (1993) definition: "An ontology is a formal, explicit specification of a shared conceptualization". Clarified further, a "conceptualization" refers to an abstract model of some "thing" in the world; "explicit" refers to the

unambiguously defined constraints and axioms; and "formal" in this context means the specification is required to be machine-readable (Studer et al. 1998; Fensel et al. 2002). Another common definition of an ontology in the context of the Semantic Web was expressed by Guarino (1997) as computer-understandable knowledge through descriptions of entities. The entities that are represented can be specific (objects) or abstract (beliefs, feelings, etc.) and the descriptions contain explicit specifications, terms and relationships, with formal definitions, axioms and restrictions, to constrain the interpretation of a concept (Guarino et al. 2009).

2.4.1.3.1. Types of Ontologies

In general, an ontology captures a domain knowledge in a generic way to provide a commonly agreed understanding of the domain that can systematically be reused and shared (Gomez-perez 1999). Because the ontology forms the basis for knowledge representations within a domain, without ontologies, or the conceptualizations that underlie knowledge, there cannot be a vocabulary for that representation. Ontologies, therefore, can span many levels of complexity, dependent on the concept to be captured and purpose for the knowledge representation. Hence, the main difference between the types of ontologies is the degree of specificity that is required (Chandrasekaran et al. 1999).

The degree of specificity can range from formal or heavyweight through to informal or lightweight when defining an ontology (Lassila and McGuinness 2001). Heavyweight ontologies are applications of formal logical definitions (e.g., Description Logic axioms), to automate conclusions, assumptions and subsumptions through classification and inference. An advantage of a more formal logical definition include finer granularity when defining concepts, which makes possible detailed explicit descriptions of entities and a deeper reasoning over Web resources. In contrast, lightweight, informal ontologies such as domain vocabularies, thesauri or taxonomies, are less complex constructs, which is advantageous when the ontology is required to be generic and flexible (Gomez-Perez et al. 2004). Specifically, the informal approach fosters simplicity for the creation and maintenance of the ontologies, and flexibility to maximise reuse (Berners-Lee, Hall, Hendler, O'Hara et al. 2006).

Specific types of ontologies have been categorised by Gomez-Perez (2004). Categories can be based on the granularity and richness of the structure (van Heijst et al. 1997; Lassila and McGuinness 2001):

- Controlled vocabularies – a finite list of terms (e.g., lists of terms used in a domain);

- Glossaries – a list of terms with meanings specified as natural language statements (i.e., definitions of the terms within the domains controlled vocabulary);
- Thesauri – a list of terms with additional semantics to define relationships between terms (i.e., synonyms, antonyms and homonyms), but without an explicit hierarchy;
- Informal and formal "is-a" hierarchies – contain additional hierarchical information, including synonymous, antonymous and hyponymous relations, that are added so reasoning engines can perform subsumption and classification tasks. A taxonomy is an example of an informal ontology type; and
- Heavyweight ontologies – contain restrictions, constraints and axioms for highly expressive explicit specifications of a concept.

The type of ontology can also be categorised by the subject and structure of the conceptualization (van Heijst et al. 1997):

- Knowledge Representation (KR) ontologies – are knowledge modelling ontologies that capture representational primitives used in a given KR paradigm (van Heijst et al. 1997; Gomez-Perez et al. 2004). Examples of KR ontologies are the formalisms used to model concepts, for example the RDF, RDFS and OWL KR ontologies;
- General ontologies – represent common sense knowledge and are reusable across many domains (van Heijst et al. 1997). The unit ontology, a part of NASA's Semantic Web for Earth and Environmental Terminology (SWEET)[11] set of ontologies, defines standard measurement units used today and is an example of a general ontology;
- Top-level or upper ontologies - describe very general concepts that are the same across all domains (Guarino 1998). Wordnet [12], is a lexical database for the English language and is an example of an informal upper ontology. In contrast, CYC[13], an example of a formal upper ontology, consists of axioms and explicit definitions to constrain ambiguity and allow for reasoning. CYC codifies, in machine-usable form, the millions of pieces of knowledge that comprise human common sense;
- Domain ontologies – provide the vocabularies, relationships and principles of a domain, and are reusable within that domain (van Heijst et al. 1997). The Gene ontology[14], which is highly applied in the bio-informatics field; is an exemplar domain ontology;

[11] http://sweet.jpl.nasa.gov/index.html
[12] http://wordnet.princeton.edu/
[13] http://www.cyc.com/
[14] http://www.geneontology.org/

- Task ontologies - describe the vocabulary of a task or activity, such as a workflow, a scheduling task, a supply chain, etc. (Guarino 1998);
- Domain-task ontologies - are task ontologies within the specific domain; and
- Application ontologies - are usually heavyweight ontologies that have been designed to achieve a particular purpose and are application dependent (Guarino 1998; van Heijst et al. 1997).

The degree of an ontology's reusability is directly relative to the degree of complexity at each level (Gomez-Perez et al. 2004). The ontology design applied in this study aimed to maximise reuse and functional usability. Therefore, a range of ontologies were created, from informal reusable ontologies (e.g., the taxonomy of reef community stock) to the formal domain, domain-specific and application ontologies and are discussed further in Chapter 3.

2.4.1.3.2. Ontologies and Data Integration

The development of semantic technologies is driven by the need to integrate data. Today, most web information is represented in natural-language but computers cannot understand and interpret its meaning. Specifically, everything on the Web is machine-readable, but not machine-understandable, because the Web was built for human consumption and not for machine consumption (Lassila 1998). Ontologies are applied to bridge the disparate data held in data silos and/or the information available on the Web with explicit definitions that enable the linking of datasets, which could not previously be connected because of the different vocabularies (Goble and De Roure 2004). Hence, the semantic technologies encompassed under the Semantic Web umbrella aim to make raw data and information "understandable" to a computer, through added contextual information, and thereby enable the computer to make intelligent decisions based on inference rules and DL (Antoniou and van Harmelen 2004).

The contextual information, supplied within the ontologies, can be applied to link concepts that are ambiguous to a computer. Ambiguous words such as homonyms require added context so the computer can distinguish between the different meanings. For example, the word "fluke" is a homonym that has many meanings, it is: a class of flatworm; a species of fish; the pair of tail fins on a whale or dolphin; a term for a stroke of luck; a flat bladelike projection on the arm of an anchor; a barb on a harpoon or arrow; a 1977 novel by James Herbert; a 1995 film based on Herbert's novel; and many more. However, to the machine, "Fluke" is five characters of eight binary digits each; it has no meaning. Only after contextual terms and information, such as "part

of" or "is a" relationships, are expressed can the computer "understand" the context in which the word is used and subsequently make automated decisions about its meaning (Köhler et al. 2006).

Ontologies link data and linking data has many benefits for data integration (Berners-Lee 2008). Once data is linked and easily parsable and understandable by the computer, and in formats such as RDF and OWL, data integration at a high, user-oriented level is achievable. Some benefits include, but are not limited to:

- The increase in productivity of research and development. For example, if datasets, metadata, etc., are more accessible through automated functions, processing can be easier and faster (Gomez-Perez et al. 2004);
- The potential to create new disciplines by linking discipline-specific languages and to create more interoperable datasets (Goble and De Roure 2004);
- Website and/or document organisation and navigation are enhanced through:
  - Support for structured, comparative and customised searches and more intelligent browsing and searching through generalisation or specialisation of the search items;
  - Sense and context "disambiguation" support, via well defined meaning to the content held on the computer or presented in Web documents;
  - More thorough capabilities in consistency checking by integrating explicit restrictions processable by the computer;
  - Controlled vocabularies offer interoperability support through equivalencies and differences in the user and application terminology;
  - Auto completion and natural language support; and
  - Support for validation and verification testing through provable inconsistencies and conditions (McGuinness 2002).

2.4.1.4. The Ontology Languages

The Semantic Web languages represent information and simultaneously make that information both syntactically and semantically interoperable across applications. RDF and OWL are KR languages for representing concepts and are employed as the main languages within the scope of the Semantic Reef Knowledge Base (KB) (Gomez-Perez et al. 2004).

2.4.1.4.1. RDF and RDFS

RDF is a simple KR model which represents resources by declaring descriptive statements called triples. Anything can be defined as a set of triples. RDF is based on the identification of resources via URIs and then describes the resources in terms of their properties and property values (Brickley et al. 2004; Powers 2003). Each component of the triple is assigned a URI and is the foundation for processing metadata because information is provided about the resource that is understandable by the computer (Powers 2003; Goble and De Roure 2002). The example in Figure 2.3 depicts an RDF triple to state "carnivores eat meat" and the predicate, in this case "eats", is a verb that describes the relationship between subject (carnivores) and object (meat). These machine-processable descriptions are uniform and standard without being inflexible or constraining. Triples can be used by the higher level semantic languages and standards to create new knowledge, make extensible searches and connect disparate data (Antoniou and van Harmelen 2008).

RDF Schemas provide a way for the RDF descriptions to be combined into a single vocabulary. They extend the RDF triple to make a more structured concept description by adding modelling primitives with fixed meanings and constraints, such as domains, relationships, subclasses and property and sub-property relations (O'Hara and Hall 2009). For example, the declaration of a domain and range restriction on a property, which confines individuals from one domain to link only with individuals of the designated range, can be processed by a reasoning engine for classification of the individuals.

RDFS provides very wide interoperability, however, it is minimalist and unable to capture a complete semantic logic because it provides only a limited number of descriptors that support

```
<triple:Clause>
        <triple:head>
          <triple:Triple>
                <triple:subject>
                        <triple:Resource rdf:about="#Carnivores"/>
                </triple:subject>
                <triple:predicate>
                        <triple:Resource rdf:about="#eat"/>
                </triple:predicate>
                <triple:object>
                        <triple:Resource rdf:about="#meat"/>
                </triple:object>
          </triple:Triple>
        </triple:head>
</triple:Clause>
```

Figure 2.3 – The statement "Carnivores eat meat" as an RDF triple statement

inference (Lacy 2005). RDF is the framework that can define vocabularies as objects, and give them structure through properties and classes. In contrast, OWL refines the vocabularies through more extensible descriptive constructs (McGuinness and Harmelen 2004).

2.4.1.4.2. OWL

The complexity of an ontology is relative to the purpose for which it is created. As described earlier, the many types of ontologies are diverse, ranging from very informal, such as a domains vocabulary, to highly formal that make possible automated classification or inference. The OWL standard adopted a range of KR models that made complexity available to suit the requirements of the ontology to be created. The strategy aimed for total expressiveness, flexibility and scalability while maintaining maximum efficiency for reasoning support. Due to these requirements, an ontology language of "one size fits all", was simply not feasible (Antoniou and van Harmelen 2008). OWL, which is W3C's recommended specification for an ontology language, has three different varieties: OWL Lite, OWL DL and OWL Full. Each of these sublanguages is an extension of its simpler predecessor (McGuinness and Harmelen 2004).

OWL Lite supports the classification of hierarchies through the available limited constraints, such as binary cardinality. Ontologies structured in OWL Lite define uncomplicated class hierarchies; for instance, a thesauri or taxonomy which can be defined by basic axioms and constraints (Smith et al. 2004). The computational burden on the reasoning and inference tools is minimised at this level due to the less complex ontology concepts. While OWL Lite provides an extended support for representing information, in many circumstances more refined descriptions are desirable and need additional language constructs (Lacy 2005).

OWL DL was designed to work with reasoning systems and was named due to the correspondence and support given to Description Logics, which is a decidable fragment of First Order Logic (FOL) (Smith et al. 2004; Lacy 2005). OWL DL is more expressive then OWL Lite as it contains the whole OWL vocabulary and fewer restrictions. In fact, OWL DL includes all OWL language constructs with only restrictions in the separation of construct type or, more precisely, a class or property cannot belong as an individual to another class. OWL DL, as a formalism for representing knowledge, supports maximum expressiveness without losing computational completeness (i.e., all statements can be computed and will finish in finite time) and decidability of reasoning systems (Baader et al. 2007). OWL DL

OWL Full is the most expressive OWL sub-language. It provides the syntactic freedom of RDF with no computational guarantees. Sometimes complete representation of an ontological

domain may be required, even though it cannot be guaranteed to be internally consistent. OWL Full is intended for use in situations where very high expressiveness is more important than being able to guarantee the decidability or computational completeness of the language. Therefore, it contains all OWL primitives, and because an arbitrary combination of those primitives with RDF and RDFS is allowed, the degree of expressive power makes it impossible to perform automated reasoning (Smith et al. 2004; O'Hara and Hall 2009).

The sublanguages that best suit the needs and the purpose of the ontology are important considerations during the ontology development. The extent of expressiveness in the constructs required would determine whether OWL Lite, OWL DL or OWL Full was the appropriate language. The choice between OWL DL and OWL full depends on the extent of meta-modelling facilities of the RDF Schema that are required, specifically, if there was a need to assert a class as an individual of another class (Antoniou and van Harmelen 2008) , which in the case of this study was not a necessity. The logical semantics of OWL DL (and Lite, which is a subset of DL) are based on DL; therefore, all inferences available in an OWL Lite or OWL DL ontology can be computed using the reasoning engine. Conversely, ontologies in OWL Full are not decidable and have insufficient application reasoning support available. Therefore, to maximise the full richness the logic systems for reasoning and inference functions, the ontology design within this project were maintained in OWL DL or Owl Lite.

2.4.1.5. The Logics - Reasoning and Rules

The logic systems implemented as part of the Semantic Web technologies are powerful contributors to the machine automation objectives of Semantic Web**.** The logic and rules layers of Figure 2.2, which enhance OWL, support certain kinds of inference, in particular FOL systems such as DL and inference rules systems. The logic systems in this project are DL, as part of the OWL ontologies, and propositional logic, for posing inference rules. Specifically, DL is applied for automated subsumption and classification and the inference systems are employed to make deductions over the individuals of the KB (O'Hara and Hall 2009).

2.4.1.5.1. Logic Systems Differentiate KR Paradigms

Three concepts that are highly relevant in the differentiation of KR paradigms are the Open World Assumption (OWA), the Closed World Assumption (CWA) and the Unique Name Assumption (UNA) (Rector et al. 2004). Data description formalisms that are based on DL, such as OWL, support the OWA but do not make the UNA or the CWA, whereas other formalisms, such as relational database systems, support the CWA and the UNA but not the OWA. The OWA means

what cannot be proven to be true is not automatically false because there is an assumption that the knowledge of the world is incomplete (Horrocks et al. 2003). More precisely, what is not explicitly stated is considered unknown, rather than wrong, and the system simply assumes the extra information required has not, as yet, been added to the KB (Rector et al. 2004). The OWA allows the informal notion that no single KB has complete knowledge, therefore the structure is flexible and organic and can be easily modified to adapt to new or additional concepts. This is advantageous in circumstances such as describing concepts of Web available content, when the full scope of information is not immediately available or concepts can change. In contrast, the Closed World Assumption (CWA), prevalent in relational databases, sees the inability to derive truth as a false response and negation as failure. In a closed world data formalism, which by nature is highly structured and non-flexible, the addition of new fields to the schematic is a non-trivial task. The UNA means different names refer to separate entities and cannot refer to the same thing. The same object in one repository may have differing terms that refer to it from other disparate repositories, thus when querying with the UNA equivalencies cannot be inferred (Rector et al. 2004; Antoniou and van Harmelen 2008).

Ontologies enable the transition from the highly structured predefined formalisms in relational database systems to a flexible, scalable unstructured data description form. To describe most worldly concepts is too complex to structure upfront, despite modern data modelling paradigms. Ontologies offer more flexibility in design, so people can add structure as required or engines can automatically create structure from unstructured content. Unexpected changes to an ontology, to add ideas, facts or concepts, is fully supported with the OWA, whereas, a gap in the knowledge would immediately imply negation of a fact in the CWA (Baader et al. 2007). Consequently, dynamic modelling of real world concepts is more flexible and open to changes due to the support for OWA. The OWA is considered implicit in RDF and OWL, as every tuple not explicitly contained in the ontology is implicitly assumed to represent a fact that is unknown, rather than false.

2.4.1.5.2. Reasoning with DL

DLs can represent concept definitions of an application domain in a structured and formally well-understood way. DLs have concepts and roles, which are referred to as “classes” and “properties” in OWL DL respectively. The concepts are defined with membership constraints or restrictions, referred to as axioms of objects, based on their properties. An axiom is a statement of a truism, or more precisely, it is a sentence or proposition that is taken for granted as true, and serves as a starting point for deducing other truths. The axioms most common in OWL are disjoint,

class, domain and range and closure axioms. Expressed in explicit terms these axioms can then be used to automatically derive classification using a reasoning engine (Baader et al. 2005; Baader et al. 2007).

A DL KB is made up of two parts each containing sets of axioms: the terminological part (the TBox) and the assertional part (the ABox). In OWL-DL the TBox represents the ontology while the ABox represents data. In general, the TBox contains axioms to describe concept roles and relations[15] with other concepts (e.g., equivalencies and hierarchies). In contrast, the ABox contains assertions about individuals and the relationships with other individuals (e.g., universal and existential quantification) (Baader et al. 2007). The Semantic Reef KB contains multiple-ontologies and multiple-data sources, and therefore, is a DL KR system containing multiple TBoxes with multiple ABoxes.

The role of DL is to infer connections and link classes. Sub-classes have a hyponymous relationship (i.e., specialisation or "kind of") with their parent super-class and the parent class has a hypernymous relationship with the sub-class (generalisation). Many latent hyponymous and hypernymous relationships are contained in a KR system. To explicitly assert every interwoven connection between all entities in an ecosystem manually would be extremely laborious for a human, but quite simple for a computer.

The automated classification processes are handled by a reasoner, or classifier. Reasoning engines are complex applications able to infer logical consequences from a set of asserted facts or axioms. They are utilised in the KR paradigm to make assumptions and subsumptions based on the context and meaning defined in the machine-understandable axioms. A number of DL reasoners are available and the three employed in the development and testing of this project were: Racer, FaCT++ and Pellet (RacerPRO 2008; FaCT++ 2008; Mindswap 2007).

2.4.1.5.3. Inference Rules with SWRL

Ontologies can include sets of inference rules from which computers can make logical conclusions and are orthogonal to DL. Inference rules, in ontologies, supply further power in the automation process as they work in conjunction with DL to enhance representation and reasoning capabilities (Hitzler and Parsia 2009). As mentioned, DL is a set of logical statements used in OWL DL to normalise classes and individuals by assumption and subsumption and makes possible negation/complement of classes, disjoint information and existential and universal quantification. In contrast, the Semantic Web Rules Language (SWRL) is a monotonic rule system that is available

[15] A *concept* and *role* in DL is referred to as a *class* and *property* in OWL, respectively.

in the Semantic Web technologies and is built on top of OWL ontologies as an extension of OWL. SWRL adheres to the open-world model, and is thus entirely in the spirit of the OWL DL language (Hitzler and Parsia 2009; Antoniou and van Harmelen 2008). SWRL manages inference using horn-like logic, which is a subset of predicate logic (FOL), and orthogonal to DL, that is, they are not reducible to each other for monotonic reasoning (Horrocks et al. 2004).

A SWRL inference rule, which is based on the RuleML format, is atom centric. The rules contain antecedents and consequences, or the body and head respectively. The antecedent (body) of the rule represents the information supplied, and required, to draw a conclusion, and the consequence (head) is the implication that is ultimately drawn (Horrocks et al. 2004; O'Connor et al. 2005). Inference rules can be applied dynamically using a rule engine such as Jess[16] and used to infer new knowledge from the existing OWL KB (Friedman-Hill 2003; Jess 2006). SWRL functionality and examples are provided in text throughout the thesis from the development through to the performance analysis of the KB.

2.4.1.6. Relevancy - The Linked Data Movement

The concept of the Semantic Web has evolved since Berners-Lee's (1990; 2002) initial introduction. The past decade has seen an escalating advancement in semantic technologies and ontology design including the research and development of the languages, standards and tools and also the diverse practical applications in the business, publishing and medical domains (W3C 2009d; Shadbolt et al. 2006; Wolstencroft et al. 2005).

Linked Data is a method to publish data on the Web and to interlink data between different data sources. Linked Data is possible with Semantic Web browsers, just as traditional Web documents are accessed using HTML browsers. However, instead of following document links between HTML pages, Semantic Web browsers enable user and software agents to navigate between different data sources by following RDF links (Berners-Lee 2007).

As online data is made more semantically aware, via the linked data movement, accessibility to data will be more easily automated. Then, the scope of the information included in a hypothesis, will be exponentially broadened. One such application in linking data is the DBpedia project, which is an effort to publish structured data in RDF that is extracted from Wikipedia (DBpedia 2009). The DBpedia project aims to provide interlinking, reuse and the extension of data sources by facilitating open availability, inference and/or advanced querying over the rich pool of Wikipedia-derived datasets. Semantic Web projects aimed at data integration and linking, such as

[16] http://www.jessrules.com/

DBpedia, will become invaluable sources of data and information for incorporation in the Semantic Reef KB.

The Semantic Web technologies are currently being implemented and/or developed by many diverse efforts. Many of these efforts focus predominantly on the knowledge found in the documentation on web pages, while others focus on data produced by a variety of instruments. Examples of document-centric Semantic Web undertakings include:

- DBpedia, publishes structured data extracted from Wikipedia (DBpedia 2009);
- The Semantically-Interlinked Online Communities (SIOC) project, which provides a vocabulary of terms and relationships to model web data spaces, such as discussion forums, weblogs, image galleries, etc. (SIOC 2009); and
- The Linking Open Data project is a Semantic Web community-led effort to create openly accessible, and interlinked, RDF Data on the Web (W3C 2009b).
- The Creative Commons (CC) organisation is dedicated to federating data and content available on the Web through open source, open access licensing. The CC uses RDF to express license and other information about works for the Semantic Web. The CC has launched a series of projects designed to support and expand the public domain (CC 2009). Projects such as the Science Commons[17] that are focused on open access of scientific data and tools for integration and reuse, or the ccLearn[18] program that is exploring the full potential of the internet to support open learning and open educational resources to minimise legal, technical, and social barriers to sharing and reuse of educational materials.

In contrast, other semantic-oriented projects focus on data that is extracted or produced by the various scientific or structural devices. One example is the Semantic Sensor Web (SSW) project that proposes to annotate sensor data with semantic metadata to provide contextual information for situational awareness. As Sheth (2008) states "the SSW proposes that sensor data be annotated with semantic metadata that will both increase interoperability and provide contextual information essential for situational knowledge." The aim is to incorporate W3C and Open Geospatial Consortium (OGC) standardisation and extend them with Semantic Web technologies to create a SSW. Thereby providing an environment for improved query and reasoning in a sensor domain (Sheth et al. 2008). A growing series of use-cases and case studies, which adopt Semantic Web

[17] http://sciencecommons.org/
[18] http://learn.creativecommons.org/

technologies for both Web-based and instrument-based data orientation, are available at the Semantic Web case studies page[19].

The development of the Semantic Web technologies is driven to improve communication by linking different terminologies and to bridge disparate data stored on the Web (a Web of data). The primary goals in the design of these technologies are to extend the interoperability of databases, integrate and federate data, provide new tools for interacting with multimedia and to allow for people and machines to work together (Hendler 2003). Accordingly, some of the most important applications for Semantic Web technology is in the e-Science and e-Research fields, in the attempt to manage the data deluge and to create new knowledge automatically (De Roure and Hendler 2004; Johnston 2004; Hall et al. 2009; O'Hara and Hall 2009). These technologies have been applied in the Semantic Reef project to automate the data analysis process and, through a different approach to hypothesis design, help to alleviate the bottlenecks caused by the growing amount of data.

2.4.2. Grid Computing

In the 1990's Foster and Kesselman (1998) proposed Grid computing or "The Grid" as a new infrastructure for distributed computing to enable advanced science and engineering. The authors defined the technology as "a computational Grid [which] is a hardware and software infrastructure that provides dependable, consistent, pervasive, and inexpensive access to high-end computational capabilities" (Foster and Kesselman 1998). Further, the authors later refined the definition to state Grid computing is concerned with "coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organisations" (Foster et al. 2001). They compared the computing Grid with the electrical grid: the Grid distributes computer and digital centric resources to the consumer, as the electric grid delivers electricity to the home.

The Grid has the potential to change how we currently utilise networked ICT resources and services, just as the Web created a paradigm shift in the way information is shared (Berners-Lee, Hall, Hendler, Shadbolt et al. 2006; Hall et al. 2009). Grid and Web services that operate on the Internet, maintain protocols for sharing the load of decentralised dataflow The Grid includes a higher level of abstraction than the Web, which is essentially about information retrieval, because it centres on adaptive resource and service utilisation (Foster et al. 2001).

The Grid brings a solid, safe, secure infrastructure that enables Grid services to control the sharing of hardware, software and data resources. The Grid services clearly define the consumers

[19] http://www.w3.org/2001/sw/sweo/public/UseCases/

and providers through predetermined sets of rules about what is shared, when it is shared and who is allowed to share (Foster 2002). The resources to be directly accessed and shared are not simple, such as files, but rather the myriad of the hardware and software services that reside on networks (Foster et al. 2001), such as computer processing power, instruments, software programs and data. Major exemplars of traditional Grid endeavours include the Open Science Grid (OSG) (Pordes et al. 2008) and the TeraGrid (Catlett 2002), which are large collaborative projects supported by the US National Science Foundation and the European EU FP7 Enabling Grids for E-science (EGEE). These initiatives are developing service Grid infrastructures to provide researchers in academia and industry connections to major computing resources (EGEE 2009).

The Grid paradigm makes possible the creation and maintenance of Virtual Organisations (VO), which are the basis for many VREs. VOs have been defined by Foster (2002a) as "Dynamic ensembles of resources, services and people that comprise of scientific or business VO's and can be small or large, short or long-lived, single or multi-institutional, and homogenous or heterogeneous" and by Czajkowski (2001) as "a large scale sharing of resources within a formal, or informal consortia of individuals and/or institutions". A VO is ultimately an on demand, dynamic gathering of people or members (i.e., requestors and providers) in a virtual marketplace to exchange commodities and/or collaborate.

The prospective users of Grids cover the full spectrum of disciplines. Some users require access to data, which is the case here, some require great amounts of computer power for the data processing, some need access to otherwise unattainable massive-scale scientific instruments and/or facilities, and some require all of the above. Grid computing has the ability to cope with the changing demands of today's scientific needs (Foster 2002). Examples of these modern research needs include collaborative visualisation of large scientific datasets, distributed HPC and the coupling of scientific instruments to remote computers and archives (Foster et al. 2002b).

2.4.2.1. Semantic Grid

Grid technologies enable the decentralised management of resources that can be simultaneously accessed from a number of geographically separate locations. In contrast, the WWW works between server and client as a one in/one out system. The Grid has the potential to create new disciplines and business models where people and organisations collaborate in ways that simply were not otherwise possible (Foster 2002; Goble and De Roure 2004). However, Grid technologies alone will not be enough to handle the changing requirements of modern research; for instance, they enable distributed access to resources such as data and storage but not the semantic integration of that data (Foster et al. 2004). The knowledge of the semantics (i.e., meaning) of the

data and operations are not expressed explicitly in machine- understandable form, only directives hard-coded into the programs (Uschold 2003). At present, the computer or computer program, such as a software agent that performs tasks on the Web or within a grid infrastructure, cannot understand information because the information has no well-defined meaning. The incorporation of contextual information to the datasets will make input computer-comprehensible so the computer can be employed to make informed decisions about that input (Goble and De Roure 2002; De Roure et al. 2003; De Roure et al. 2005).

The Semantic Grid is an extension of the Grid and applies semantically rich information to current Grid resources to create a more intelligent Grid service (Goble and Bechhofer 2005). Grid services are interoperable and uniform due to the Open Grid Services Architecture (OGSA) (Foster et al. 2002b; Tuecke 2003). A proposal for a Semantic Grid reference architecture, Semantic-OGSA (S-OGSA), has been developed and defines a model that extends the OGSA via lightweight mechanisms that incorporate both semantic and knowledge services (Corcho et al. 2006).

Many research domains have benefited from the data integration and problem solving synergies when Grid technologies are combined with semantic technologies (De Roure and Hendler 2004). Example Semantic Grid initiatives are the European OntoGrid project (Goble and Bechhofer 2005), the GEOscience Network (GEON) in the US (Ludäscher et al. 2007) and the CombeChem project in the UK (Taylor, Essex et al. 2006; Frey et al. 2004). The OntoGrid project was an eight-partner European undertaking that investigated fundamental standards of Semantic Grid and initialised the S-OGSA. GEON is a geo-informatics research program which explores the complex dynamics of Earth systems by developing a cyberinfrastructure to interlink and share multidisciplinary data sets for integration, analysis and 4-D modelling (Nambiar et al. 2006). The CombeChem project aims to create a "Smart Laboratory" for the chemistry discipline. It takes a holistic approach to the scholarly knowledge cycle and begins with the creation of the chemical data, such as a new crystal compound. The data is then synthesised, annotated and organised simultaneously in a complete end-to-end workflow from the emergence of the new crystal at the laboratory bench to the final published analysis (Frey et al. 2002). These three varied e-Research projects exemplify the synergies from the combination of the complementary technologies such as Semantic Web, Grid computing and scientific workflows.

2.4.3. Scientific Workflows

In many domains, the nature of research is changing and the scientist's equipment is no longer just the experimental apparatus for *in vitro* or *in situ* experiments. Rather many experiments

are now digitised and performed predominantly *in silico*, such as simulations, visualisations, data mining and analyses (De Roure and Goble 2009). *In silico* experimentation conducted and controlled through workflows is now common practice (Ludäscher et al. 2006). Ludäscher (2006) considers: "Scientific workflows [as] a flexible tool for accessing scientific data (streaming sensor data, medical and satellite images, simulation output, observational data, etc.) and executing complex analysis on the retrieved data". The growth in multi-disciplinary domains such as ecoinformatics, bioinformatics, geoinformatics, etc., shows how common this revolution in digitised scientific workflows has become (De Roure et al. 2004).

Scientific workflows are employed to control *in silico* analyses. Each step in a workflow identifies the data flow and how processes, executions and computations are ordered and subsequently run (Gil et al. 2007). The research process is a spectrum which covers the full range from the initial importation and extrapolation of data to the experiment and through to the publication stage and because workflows can organise and control these processes, they have now become first class members in the experimental procedure (De Roure et al. 2007; De Roure et al. 2009). These scientific workflows are flexible and can incorporate impromptu changes to the procedure, which a scientist may wish to implement to adapt to a changing condition or modification of an experiment. The scientist has full control to design, execute, monitor, re-run and also communicate the outcome of the workflow according to need (Altintas et al. 2004).

Prime examples in the implementation of scientific workflows are the Science Environment for Ecological Knowledge (SEEK) project (SEEK 2009) and the myExperiment project. Previously, the large and disparate ecological and biodiversity data has been impossible to coordinate into one workflow. The SEEK system , however, can streamline data acquisition and archive tasks through data integration, transformation, analysis, and synthesisation (Michener et al. 2005). myExperiment is a VRE for the social curation and sharing of scientific research objects, such as research investigative designs, questions, results, publications, and in particular, scientific workflows and *in silico* experiments (De Roure et al. 2009; Goble and De Roure 2007).

Two of the most important functions for modern research strategies to evolve are the ability to automate processes and the ability to reproduce these processes at any time. Digital tools are required to be adaptive and flexible as research strategies simultaneously change and evolve (Goble et al. 2006; De Roure and Goble 2009). Software systems such as Kepler (2004), Taverna (2004), Triana (2007), to name a few, are tools that allow scientists to capture scientific workflows (Taylor, Deelman et al. 2006). Many of the modern research requirements are made possible through these systems by:

- The orchestration of numerous scientific instruments in a workflow;
- The support for experimental techniques that can be systematically merged in an ad hoc fashion, dependent on the goal;
- The support for collaborations that can span multiple organisations and allow data and information resources to be shared and reused; or
- The ability to work through massive amounts of data methodically and efficiently (De Roure et al. 2009).

Workflows make possible new strategies while scientists conduct experiments and modern researchers are now adopting or merging other domain's techniques in unique and new ways (Gil et al. 2007).

The choice of Kepler as the workflow system was motivated predominantly due to the flexibility in workflow design and manipulation. Like other contemporary open-source workflow systems such as Triana and Taverna, Kepler has an active support network and developer community. However, as shown in a taxonomic study of workflow systems by Yu (2005), Kepler is a user directed system that supports flexible data movement methods. These methods include: A centralised approach where data is transferred between resources via a central point; a mediated approach where the locations of the data are managed by a distributed data management system: and a peer-to-peer approach where data is transferred between processing resources (Yu and Buyya 2005). The flexible data movement supported by Kepler workflows enables access to a diverse range of data resources, such as the distributed data repositories and streaming sensor data required to populate the ontologies within the KB.

2.4.3.1. The Workflows for this Study

Kepler is an open-source scientific workflow system and is the tool chosen for the implementation of the Semantic Reef architecture. Kepler combines high-level design with execution and runtime interaction and can connect to both local or remote data and service invocation (Altintas et al. 2004). Kepler's functionality is described by Ludäscher (2006) as a platform which enables scientists to design scientific workflows and execute them efficiently. One such functionality in Kepler is the emerging Grid-based approaches that provide access to distributed resources such as data and computational services, while hiding the underlying complexity of the Grid technologies. The Kepler system supports the automation of low-level data processing tasks so the focus can remain on the scientific questions of interest.

The workflows that Kepler produce can be implemented in cross platform environments, provide documentation and visualisation of the processes and bring the power of distributed databases, computational Grid resources and applications to the desktop (Ludäscher et al. 2006). Each workflow step is represented by "actors," which are individual processing components that can be manipulated through a "drag and drop" method into a workflow, via Kepler's visual interface. The actors are then connected and organised according to the data flow, and the dependencies among them, to form the workflow (Altintas et al. 2004).

This ability to create varied workflows supports the need for flexibility and reuse that is required by the Semantic Reef architecture. The flexibility makes possible the formulation of separate hypotheses that may require different manipulations of data-type and data flow.

The combination of semantic technologies and scientific workflows are rare (Gil et al. 2007). In a recent publication on workflow implementation, Gil (2007) ascertained: "There's relevant work in related fields of computer science, such as, refinement calculi, model-driven architecture, semantic modelling but researchers haven't applied these techniques widely to scientific workflows". Gil was referring to the gap in the current methods for data processing, specifically in managing the data deluge. The Semantic Reef Project is a platform which incorporates both semantic modelling and scientific workflows for researchers to combine disjoint data into a single KB and to pose questions of the data.

The workflows use Web services to gain access to the data, then manipulate the data formats and pass the data on to the KB for inference or reasoning. Each step in the workflow specifies a process or computation to be executed. An example of the Semantic Reef project workflow is the use of a Web service "actor" to access sensor data from the Australian Institute of Marine Science (AIMS) (GBROOS 2008). The remotely served data is then passed to "conversion actors", such as XPATH and XSLT actors, to transform the data or extrapolate only the information required. "Computational actors" make any necessary calculations and, finally, the workflow automatically opens the KB and populates the ontologies with the relevant information. Rules or propositions are then applied to infer new knowledge or hypothesise for observational analysis.

2.5. Current Projects with a Similar Architectural Mix

2.5.1. SEEK

The Science Environment for Ecological Knowledge (SEEK) project is a large scale eco-informatics project and is a National Science Foundation (NSF) funded initiative. As previously

mentioned, SEEK is a system designed to support data acquisition and management of ecological and biodiversity data. The project encompasses many cyberinfrastructure tools that are necessary to integrate complex ecological data and enable rapid development and reuse of complex scientific analyses (Michener et al. 2005). The SEEK (2009) vision is ”to build a cyberinfrastructure which creates fundamental improvements in how researchers can:

- Gain global access to ecological data and information,
- Locate and utilise distributed computational services, and
- Exercise powerful new methods for capturing, reproducing, and analysing data by extending ecological and biodiversity analysis and research capabilities.”

SEEK encompasses three integrated systems: a Grid computing infrastructure for data storage, sharing and access; a semantic mediation system that reasons over data to determine whether it is relevant to a designated workflow; and a modelling system, for use by ecologists to design, modify and incorporate analyses when composing new workflows (SEEK 2009). The primary goal of the SEEK project is the production of an efficient tool for ecologists to capture, organise and search for data, and apply analytical processes from their desk-tops.

The Semantic Reef project can benefit from the resources made available through the SEEK facilities such as the data sources and the semantic mediation system. For example, a hypothetical proposition that is run in the Semantic Reef system can adopt the ecological data, which are available via the SEEK EarthGrid portal[20], as resources. The semantic mediation system forms a middleware component between the analytical workflow system and the data and metadata sources available in EcoGrid. The ontology-based services, provided by the semantic mediation layer, support the Kepler workflow system in data discovery and integration and offer a knowledge-based query system for the integration of disparate data resources.

Also available through SEEK, for use by systems such as the Semantic Reef, are a range of top-level formal and informal ecological ontologies. These external ontologies can be mapped to the KB because ontology design supports interoperability, scalability and reuse and enables mapping capabilities for both internal and external ontologies. Once imported, the ontologies can be modified or added to, depending on the purpose of the system. The ontologies cover unit and measurement systems and temporal/spatial concepts, among others, and can be imported to the Semantic Reef KB and adapted to suit a purpose (e.g., domain specific terms, parameters, etc.).

[20] http://ecogrid.ecoinformatics.org/ecogrid/

2.5.2. Semantic Sensor Web

The Semantic Sensor Web (SSW) project, mentioned earlier, aims to provide an environment for enhanced query and reasoning within a sensor domain to alleviate the strain of the data deluge. The SSW is an initiative of the Kno-e-sis Centre, Wright State University Ohio, US and proposes to annotate sensor data with semantic metadata to increase interoperability of that data (Kno.e.sis.Centre 2008) Thus, by annotating sensor data with spatial, temporal, and thematic semantic metadata, the SSW can provide enhanced descriptions and information essential for data discovery and analysis (Henson et al. 2009). This proposed technique builds on current standardisation efforts within the W3C and Open Geospatial Consortium (OGC) by extending them with Semantic Web technologies (Sheth et al. 2008; Kno.e.sis.Centre 2008).

The SSW applies complex queries about weather data collected from the urban Geographic Information System (GIS) systems and weather services in an exemplar use-case. A prime motivation of the SSW is to merge the data gathering instruments (e.g., remote sensors, video and other cameras devices, etc.) with the collection and analysis process. This merger is important because there is currently a lack of integration and communication between multi-layer sensor nodes, such as high-level and low-level sensor networks. The information, once integrated to the SSW, is valuable to many query or inference applications, such as traffic control, weather condition alerts, crime detection, sensor quality and fault control of traffic devices. Data, such as air, surface, subsurface, and dew point temperatures, as well as wind speed, wind direction, and precipitation are collected and then assimilated. The data can then be queried, reasoned over and/or have inference

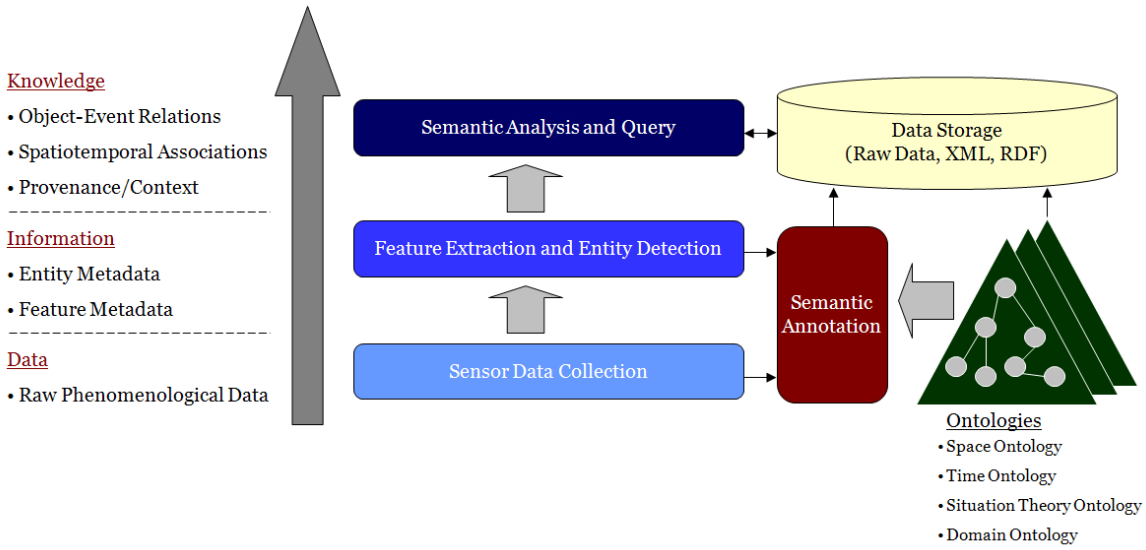


Figure 2.4 – The Sensor Semantic Web Architecture (Sheth 2008).

rules applied, including rules to automate alerts to possible weather and/or road conditions (e.g., potentially icy traffic conditions or blizzard conditions) (Sheth et al. 2008).

The SSW focuses predominantly on the annotation and quality control of sensor data. The focus of the SSW is to explore higher semantic functionality within the sensor technology standards and proposes new additions to the current sensor standards. The proposal includes the addition of semantic annotation to the sensor layers as metadata to sensor data for access to sensor data streams. When the data is relevant to marine research the data in the storage level of the SSW architecture (Figure 2.4) will would be a valuable source of quality assured sensed data for import to the Semantic Reef system.

2.5.3. NOAA's ICON/CREWS

Integrated Coral Observing Network (ICON) is a coral reef monitoring program providing early warnings and long-term monitoring of key coral reefs, both domestic and international, and uses both satellite and remotely sensed data. The National Oceanic and Atmospheric Administration's (NOAA) Coral Health and Monitoring Program (CHAMP) (2006) provides services and information sources to help improve and sustain coral reef health throughout the world. The ICON program is a CHAMP initiative and is a collaborative effort between two organisations within NOAA: Oceanic and Atmospheric Research (OAR), and the National Environmental Satellite, Data and Information Service (NESDIS) (NOAA-ICON/CREWS 2008). Currently, the tools developed by ICON combine current data streams and historical data to provide reef management and researcher support and information. The ICON architecture is described by Hendee (2008) as: "a series of artificial AI techniques to produce near real-time data-driven models of how organisms or events are influenced by meteorological and oceanographic stimuli acting singly and synergistically". The models that ICON produces are ecological forecasts which aim to predict the impacts of physical, chemical, biological, and human-induced change on ecosystems to predict, for instance, coral bleaching events.

The ICON model is an inference application that applies rules for both deductive and abductive reasoning. The architecture employs a heuristic modelling approach, which is a belief-based problem-solving technique, and the Stimulus/Response Index (S/RI) as the basis for the rules. The S/RI is a numerical measure to determine the response by organisms and ecosystems to impact pressures, such as temperature and light. To determine the S/RI, both concrete, historical fact as well as subjective data (the beliefs, conjectures and knowledge of the expert community), are taken into account. To date, the ICON model has successfully modelled coral bleaching response to coral

stressors and also represents an important advance on simply presenting raw data streams to reef managers (Hendee and Berkelmans 2003; Hendee et al. 2008). The tools (e.g., the specialised G2 Server) and rules are used to reduce data to simplify the ecological forecasts and is a prime exemplar for data integration.

2.5.4. OntoGrid – QUARC

The OntoGrid project endeavours to examine fundamental Semantic Grid concepts that bridge the knowledge-based systems community and the Grid community. The OntoGrid project aimed to bring together knowledge services, such as ontology services, metadata stores and reasoning engines, with Grid services such as workflow management, VO formation, resources brokering and data integration (Goble and Bechhofer 2005).

The Quality Analysis of Satellite Missions (QUARC) was a case study within the OntoGrid project and focused on satellite data management. QUARC demonstrated a practical exemplar of a Semantic Grid architecture that applied quality analysis, query processing and transference of data to different autonomous systems in satellite missions (Sanchez-Gestido et al. 2006). Built on a Grid system and initiating the S-OGSA services, the QUARC experiment exposed the metadata, which was in RDF, from data products obtained in satellite missions (e.g., complex imagery devices). Semantic queries with query languages, such as SPARQL, were then applied (OntoGrid 2007). The system exploits the data to query the provenance of information in satellite instrument planning in order to improve the overall quality of this data. QUARC returns reports and plots that are designed to assist when an instrument or the whole system begins to malfunction, for instance when incorrect anomalies that may occur in data product generation or data circulation are detected (Wright et al. 2008; Sanchez-Gestido et al. 2006).

2.5.5. Health-e-Waterways

The Health-e-Waterways[21] project aims to develop a cyber-infrastructure to assist management in the decision-making challenges of the Australian State of Queensland's water supply. This project enables a collaborative integration and analysis of high quality data and information about water, such as the tracking of water movement, consumption and quality. Due to climate change, urban development and population growth, there is a need for technological solutions that allow scientists, urban planners and policy makers to track water through the entire supply process (Alabri et al. 2009).

[21] http://www.healthywaterways.org/Home.aspx

The Health-e-Waterways architecture is a combination of Semantic Web technologies, scientific data servers, Web services, GIS visualisation interfaces and scientific workflows. A select group of quality assured datasets and models will be integrated and shared through a combined water information management system and a web portal. The first implementation for the Health-e-Waterway's architecture is a more efficient approach to the production of the annual Ecosystem Health Monitoring program (EHMP) report cards[22]. These reports are used by Australian politicians and planners, local councils, universities and research institutions as a valuable source of information for making decisions about land use, water quality, allocations and investments in water recycling plants.

Once Queensland's water information is united and is made universally available it will be a valuable source of data for the Semantic Reef system. The salinity, turbidity and hydrology of reef systems in offshore areas are affected by coastal waterways (AIMS 2008). This combination of complex data will be available from the Health-e-Waterways system and can be applied in hypotheses. Notably, data from the Health-e-Waterways data server will be obtainable via Web services and can be imported to the Semantic Reef KB for use in observational hypotheses and predictions.

2.6. The Marine Science Domain

Globally, marine ecosystems face many pressures from both natural and human-induced stresses, and scientific insight and global management coordination is required to overcome these threats. According to the Australian Institute for Marine Science (AIMS) dangers to Australia's coral reefs fall into three categories (AIMS 2007b):

- Natural stresses of which corals have evolved to cope with for millions of years;
- Direct anthropogenic pressures that include sediment and nutrient pollution from land run-offs, fishing practices that damage and overexploit fish populations and the engineering and modification of shorelines; and
- Global climate change and variability.

Studies of the manifestations of global climate change in coral reefs, such as increased coral bleaching and coral disease, have shown that many of these threats are closely linked and exacerbate each other (AIMS 2008).

[22] http://www.ehmp.org/annual_report_cards.html

2.6.1. Example Hypothesis - Coral Bleaching Alert

The impact of climate change on the Great Barrier Reef (GBR) endangers this iconic wonder. In fact, a major contributor to the coral bleaching phenomena is climate change induced warming of ocean temperatures (Hughes et al. 2003). Corals live in a symbiotic relationship with single-celled dinoflagellates called zooxanthellae that live in coral tissue at extremely high densities. Coral bleaching results from a stress condition in corals that induce a breakdown of the symbiotic relationship between coral and zooxanthellae.

The survival of both organisms relies on this symbiosis. Zooxanthellae reside in every cell of the coral animal's tissue and in exchange provide energy-rich sugars via photosynthesis, which is a major food source for the coral. Notably, the photosynthetic pigments from the algae give corals their brilliant colours and reef corals are very sensitive to Sea Surface Temperatures (SST) outside their normal range. Stress factors in many coral species are triggered when the temperature is sufficiently elevated for the coral to expel the zooxanthellae. When the zooxanthellae are removed, the white skeleton of the coral is exposed which causes the bleached white appearance (Jones et al. 1998).

Coral bleaching events have been largely attributed to anomalously high temperatures but studies have shown they can also be caused by other factors, or a combination of them. Such factors include, but are possibly not limited to, low-salinity, high-light intensity, pollutants, exposure, pH and even sedimentation (Hughes et al. 2003). Current research into the coral bleaching phenomenon includes investigations to find the tipping point for coral death given different combinations of the ecological factors and stressors. A tool such as the Semantic Reef system that can pose hypotheses over disconnected datasets and automate inferences about the tipping point would assist these hypothetically-driven research efforts.

2.6.2. The Data Problem

The synthesis of acquired knowledge and large multi-disciplinary data sets are necessary to find solutions to the range of problems currently facing coral reefs. Notably, many of the most influential papers in coral reef science of the past few years have been "synthesis" papers, aggregating long-term observations into new hypotheses and conclusions. One well known example of this new class of science is the recent Inter-governmental Panel on Climate Change reports (IPCC) (IPCC 2007). The IPCC reports are the product of many scientists working together in an international, intergovernmental, scientific federation, to foster an efficient research paradigm on climate change. Through activities such as the IPCC, future research will be conducted by

finding correlations in data and through observational hypothetical studies in the ecological domain.

The collection of oceanographic and marine data is a complex and expensive process that requires significant scientific and technical infrastructure (Huddlestone-Holmes et al. 2007). Decisions about where and when to collect data to ensure the most efficient, cost effective, best use - and ultimate reuse - of all the data collected, is imperative.

The marine biology domain requires hypothesis-driven research tools and problem-solving methods for efficient investigation of the disparate data streams and data sources (Hey and Trefethen 2003a). Environmental sensor networks are deployed to gather data in real-time across widely distributed areas for applications such as environmental and seismic monitoring. The Integrated Marine Observing System's (IMOS) Great Barrier Reef Oceans Observing System (GBROOS) is one such infrastructure for remote monitoring of chosen sites on the GBR via sensor networks (IMOS 2008). GBROOS and similar sensor network deployments are expected, when fully implemented, to produce vast amounts of real-time streaming data from a variety of domain-specific sensors including meteorological, chemical and biological. The vast amount of sensed data is necessary for environmental management, particularly in the production of new information that will expand knowledge and understanding of the affects of climate change.

2.7. The Semantic Reef Project

The Semantic Reef Project is a platform that consists of scientific workflows and a semantic KB so researchers can combine and question scientific data. The scientific workflows retrieve remote sensor data and data available via the Web and integrate the data into the existing KB for further synthesis and analysis. Throughout this process, the automated workflow performs any necessary calculations, reformats the data and then routes it into the KB. The semantic KB consists of a hierarchy of ontologies to describe a coral reef ecosystem that can enable ontology-based data integration. Once the ontologies have been populated by the workflow, the data can be reasoned over and inferences can be made. For example, a domain expert, either a marine scientist of reef manager, can query the KB to extract information of interest, pose observational hypothesises or as an alert system by inferring events (refer to Figure 2.5). The Semantic Reef model is a research case study that combines semantic technologies, scientific workflows, FOL and propositional logic systems. The model represents a proof of concept exemplar of future methods for managing rich data sources in more productive ways.

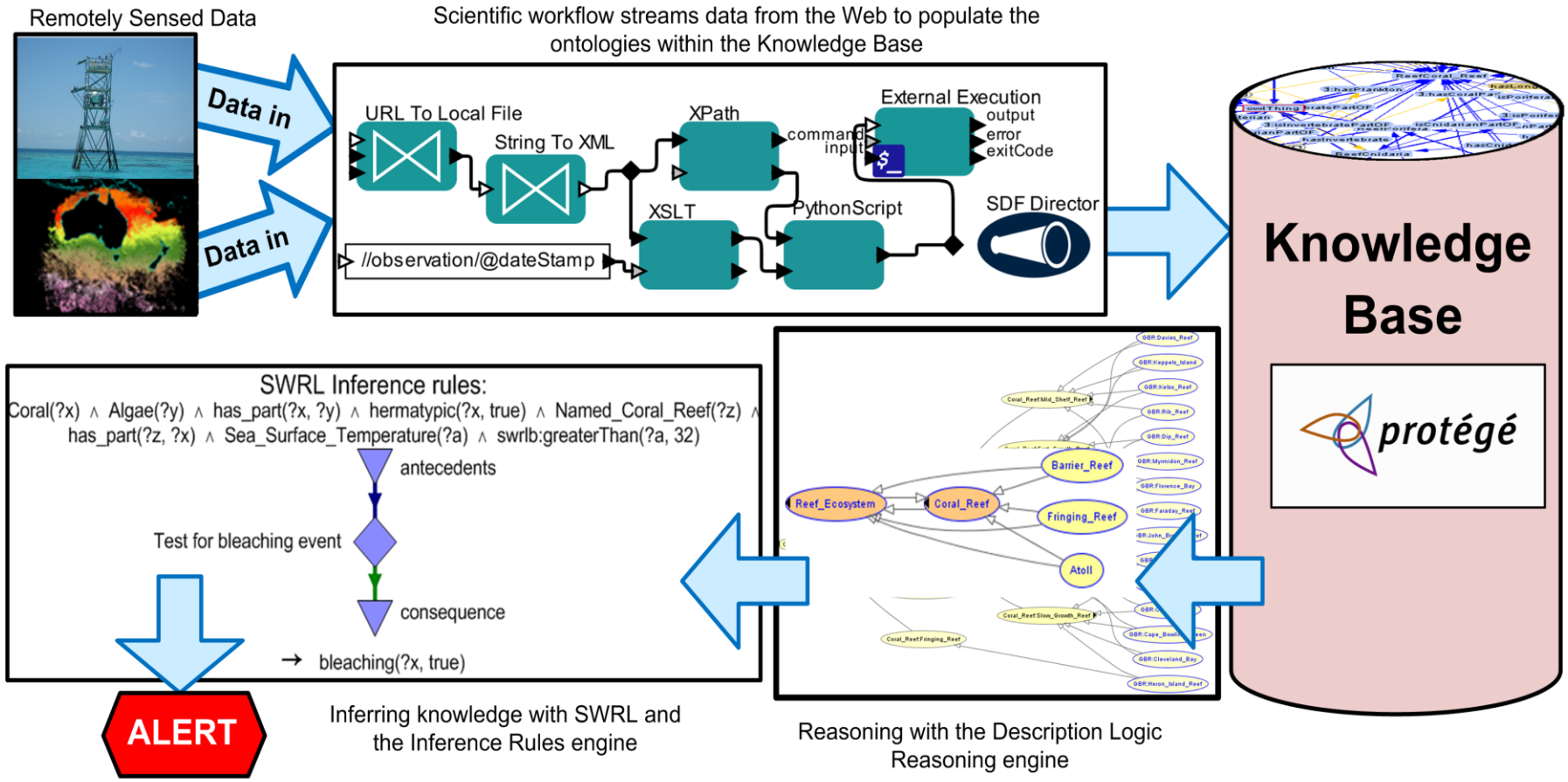


Figure 2.5 - An example Semantic Reef Workflow that results in a bleach alert.

Assistance to streamline the data processing and analysis phase of research is required by modern researchers. Ideally, scientists are best left to concentrate on the research studies, not on the ICT tools they use to process the data. The need for workable solutions to automate the data analysis steps has arisen due to scaling-out data sources.

The Semantic Reef project aims to balance the inevitable growth in data, in particular remotely sensed data, with the capacity of the individual researcher to create new knowledge. The Semantic Reef model assimilates a number of e-Research technologies that each offer solutions to help resolve the data processing and data management bottlenecks. The automation of data analysis processes should free the scientist's time to work on research and not be taxed needlessly with data access and/or reformatting. For example, the automatic integration of disconnected data sources, with different formats (e.g., spreadsheet data, annotation data, live streaming binary data, etc.), for inclusion in a hypothesis, would enrich the hypotheses tested with additional facts and also allow the researcher to continue with other aspects of the study.

The system consists of a workflow driven KB that is a tool to test hypotheses and forward-chaining probabilities within a semantic environment. The KB contains a hierarchical set of reusable and usable ontologies, from informal to formal, and the major strategies in the ontology design were aimed to achieve flexibility and reusability. The level of granularity required for a line of enquiry, or the depth of the analysis, is flexible and scalable and independent of reef location, reef type or its community composition. The system's reusable modular design purposefully separates the ontological functionality and relationships, which are emulated in the integrated logic systems, from the instance data. Once the disparate data is coupled to the ecological and environmental ontologies described herein, queries and propositional inference can be executed.

The outcomes of the hypotheses can provide information for application with different agendas. The information can provide marine biologists with new knowledge or marine park managers with reports to assist in decision-making about reef issues. For example, hypotheses for *in situ* observation can be posed to find the combination of contributing factors that make up the tipping point for coral death by bleaching. Alternatively, alerts can be inferred for unusual domain-specific events, such as coral spawning or coral bleaching, to assist marine park managers (refer to Figure 2.5).

Further, the Semantic Reef system has the potential to expose data gaps during the hypothesis process, because all relevant data is required to fill the premises of the propositions to infer an outcome for either an alert or a hypothesis. Explicitly, if there is no specific data of that

location, or the desired environmental factor is not currently monitored, it would be impossible to run the hypotheses or to conclude a legitimate outcome.

Notably, data gathering endeavours are costly in the coral reef research domain, and thus exposing gaps in the available data is important for making decisions about data acquisition. These decisions include, but are not limited to, the location to deploy the instruments, what elements are to be measured and what metrics will be used. To exemplify the problem, sensor networks in the oceanic environment include costs of deployment and maintenance as well as decisions about the sensor type, what it will measure and where to locate the device. The maintenance is extremely expensive due to the remote locations and harsh environments because the location is reachable only via boat. Also, corrosion occurs because salt water is not an ideal environment for an integrated circuit.

Due to the extreme conditions, constant maintenance is required in cleaning, re-calibrating and/or exchanging the sensors and each requires travel to and fro, which subsequently adds to the cost of data collection (Rajasegarar et al. 2008; Huddlestone-Holmes et al. 2007). Hence, when making decisions or prioritising the most efficient and cost effective deployment strategy for data acquisition, managers of funding resources would benefit from predetermined knowledge of gaps in the data.

2.7.1. A Comparison of Architectures

This project is an exemplar solution focused on one domain, coral reefs, and is not a conventional implementation of a semantic-orientated eco-informatics architecture. Instead it is a "stretch" implementation, designed to test the limits and capacity of the Semantic Web model in real-world problem solutions.

A matrix to contrast the architectural characteristics of the Semantic Reef project with the other eco-informatics exemplars, described in section 4, is shown in Appendix A. The SEEK infrastructure, the SSW, OntoGrid's QUARC use-case, the ICON program and the Health-e-Waterways project, are compared. Although exploring new ways to resolve, or at least quell, the deluge is a motivation common to all of the projects, each have different agendas and strategies in doing so. The contrast of similarities and differences between the projects and the different strategies in methodology, approach and the mixture of technologies employed to achieve their distinct goals, is illustrated in the matrix (Appendix A). The four categorical topics explored are:

1) The data-type and data integration limitations and constraints to a scalable general purpose platform;
2) Whether the model is a query system or a hypothesis system;
3) Whether workflows have been enlisted to assist in automation and control; and
4) The level of complexity of the incorporated semantic technologies, which is illustrated in Figure 2.6.

2.7.1.1. The Data Sources and Data Integration

The level of quality assurance required for the types of data incorporated in each project affects the strategies, methodologies and/or scalability. The limitations, flexibility and scope of the projects depend on the data sources, the level of quality control of that data and the methods or use of ontology-based data integration. The dependencies include whether the data must be from a quality assured source or completely open source; whether the project can only use data from a preset number of sources (data silos, distributed data, etc.); and/or whether the data has temporal limitations (historical data versus real-time streamed data). The Semantic Reef architecture is a scalable general purpose ontology-based model that permits any digitalised data from any openly available source.

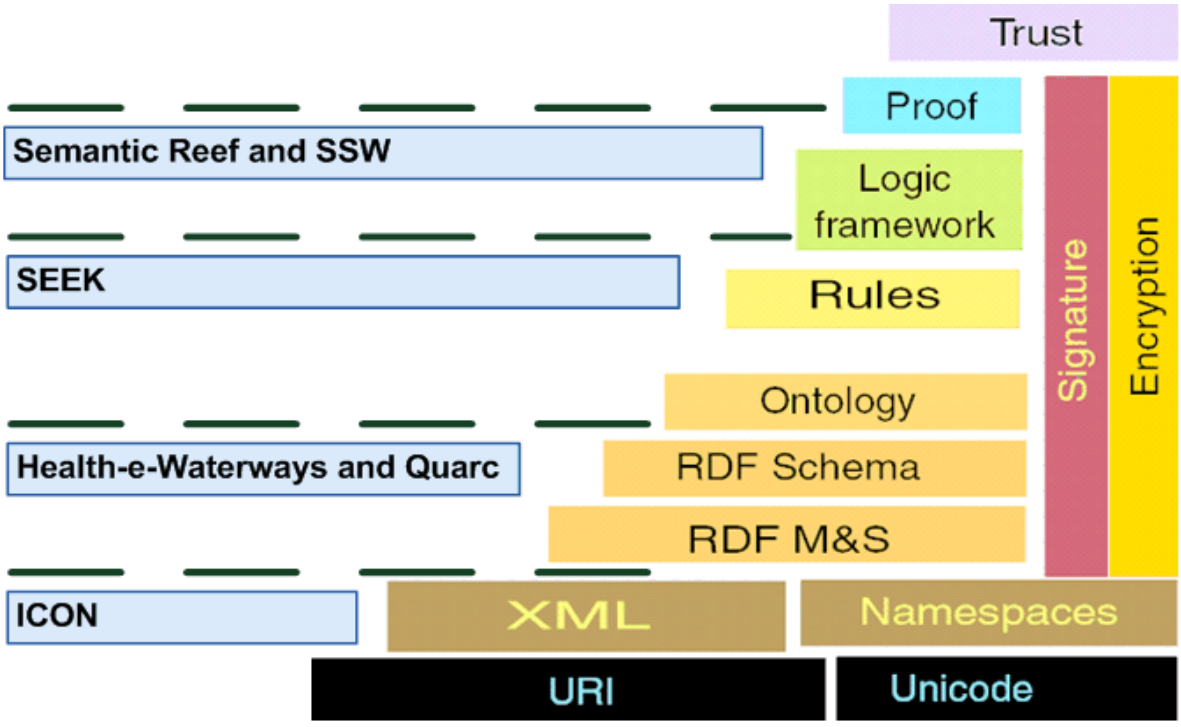


Figure 2.6 – The level of Semantic Technologies employed by the projects

Besides the Semantic Reef project, SEEK is the only other eco-informatics project surveyed that is designed to incorporate open data. SEEK is a support infrastructure with a holistic view of the eco-informatics domain and permits scientists to structure their own experiments. In contrast, the Semantic Reef model is an atomistic application that focuses predominantly in the subset of coral reef ecosystems. SEEK offers beneficial resources that include data sources and the semantic mediation system. The semantic mediation layer provides ontology-based data integration services that are supported by the Kepler workflow system in data discovery and integration and offer a knowledge-based query system for the integration of disparate data resources.

All other projects necessitate quality assured data from set data sources and do not employ extensive ontology-based data integration methods. Specifically, the QUARC project operates with designated satellite datasets and the quality control process is its main function. The ICON, Health-e-Waterways and SSW projects utilise temporal, geospatial and environmental data from controlled data gathering instruments, distributed sensor networks and/or satellite systems, after quality assurance procedures have been applied. Further, in the case of Health-e-Waterways, a number of disparate data silos are also integrated such as water usage data by demographics. The Semantic Reef project can incorporate any relevant openly available data source and the strategy for quality assurance is a component of the methodology of the researcher. More specifically, the quality of data is the decision of the specific researcher and the level of validity and credibility of the data sources will be incorporated in the methodologies of the individual hypothesis.

2.7.1.2. A Query System or Hypothesis System

The technologies each project employs define whether it is a query system, a hypothesis system, or both. A query system extracts data, information or knowledge via a data manipulation language such as Structured Query Language (SQL) with well defined look-up or keyword searches. In contrast, a hypothesis system derives or infers knowledge via explicitly asserted facts (axioms).

Semantic queries, for searching and retrieval of instances in RDF, are incorporated in the Semantic Reef model through support for SPARQL. The capability to pose propositional inference through SWRL rules, to query and infer over OWL instances, makes observational hypotheses possible in the Semantic Reef system. Both the query and hypothesis systems are functionalities built into the Protégé ontology authoring tool, which was employed to create the Semantic Reef KB (Protégé 2009). In comparison, although all the projects in the study have database query support of some kind, such as Query-By-Example, SQL, or other proprietary forms, ICON and SEEK were the only ones that did not implement semantic querying using SPARQL. Furthermore, the only

other project that offered support for hypothetical queries using propositional logics is the SSW which, like the Semantic Reef, utilises all layers of the semantic technologies (Figure 2.6).

2.7.1.3. Workflow Support

Workflows are an important tool, particularly in e-Research for the automation, organisation and streamlining of processes. As previously mentioned in section 2.3, there are gaps in the current methods for data processing in managing the data deluge. Related fields such as model-driven architectures and semantic modelling are developing possible solutions to streamline data analysis; however, scientific workflows have not been widely applied to these techniques (Gil et al. 2007). Tools are required that let domain scientists effectively harness the functionality of an e-Research infrastructure without the need to become computer scientists themselves. Currently, the most common tools that enable this functionality are those that have adopted portals and workflow environments (Hey and Trefethen 2005; Ludäscher et al. 2007; Gil et al. 2007). Of the projects surveyed, the Semantic Reef, SEEK, QUARC and Health-e-Waterways employ workflows within their architecture to automate processes, initiate data access, integration and processing tasks. Conversely, the SSW and ICON do not.

2.7.1.4. The Application of Semantic Web Technologies

The level of semantic technologies incorporated into the projects differentiates the various architectures of each project. A version of the Semantic Web "layer cake" (Berners-Lee 2003; Fensel et al. 2002) is presented in Figure 2.6 and has been expanded to depict graphically the semantic level of each project. ICON employs the lower layers of the cake by default, as they are the basis of the services and standards for access and deployment via the Web. Notably, because the semantic layers begin at the RDF level, ICON is stated as not employing semantic technologies. Alternatively, the ICON architecture enlists heuristic methods to predict coral reef environmental events, which is an appropriate AI technique for its purpose although limited for other logic based propositions.

The Health-e-Waterways and QUARC projects implement RDF triplestores for semantic queries and reasoning, but do not initiate inference or logics at the higher ontology design levels. More specifically, they use RDF taxonomies to describe database schemas and informal OWL ontologies to define satellite instrument concepts for quality control and fault detection of the data and instruments.

The SEEK initiative incorporates the semantic mediation layer as a component to its architecture. The mediation layer reasons over data to determine data relevancy and analytical

components for automatic transformation and use in a selected workflow. SEEK also applies the higher levels, such as description logics, within some of the environmental ontologies. The ontologies are maintained as a repository for public access and use (e.g., the food web or biodiversity ontologies) and are part of the infrastructure offered by the SEEK program. Rules based inference is not supported as a component of the SEEK implementation.

The Semantic Reef and the SSW are the only projects to incorporate all the logic and inference systems available in the Semantic Web stack. The agenda of the Semantic Reef project was to explore the possible benefits these technologies offer to hypothesis-driven research in the marine science domain. In contrast, the SSW is not focused on a single domain use-case unlike the Semantic Reef system. Rather, the SSW agenda concentrates predominantly on higher semantic functionality within the sensor technology standards. Specifically, the SSW proponents propose semantic annotation to be added to sensor layers as metadata to sensor data for access to sensor data streams from any given application. The implementations of the standards proposed by the SSW agenda for the annotation and quality assurance of sensed data will produce valuable resources for the Semantic Reef system in future hypotheses.

In conclusion, parallels between the projects presented in comparison to the Semantic Reef project are apparent. Accordingly, the Semantic Reef project would be enhanced by leveraging the products or standards these larger endeavours create. For example, the products and data available through ICON, Health-e-Waterways and SEEK have either been utilised as data sources or will be in future work. Additionally, projects such as SSW and OntoGrid work toward the enhancement of current standards and technologies by incorporating semantic concepts in proposals for data and information management specifications.

2.8. Summary

E-Research and the different data integration techniques, methods and strategies, have been widely reviewed. The enabling technologies that are included in the e-Research paradigm can increase the productivity, efficiency and scope of the research process. A discussion of the data deluge articulated the data management problems faced by modern researchers, in particular, the bottlenecks that are arising in the data analysis phase. The key technologies that offer solutions to the management and analysis of data in modern research were discussed. These technologies included, Semantic Web, Grid computing and scientific workflow tools.

The Semantic Reef project, which is the focus of this thesis and the domain of interest in its implementation (i.e., coral reef ecosystems) was introduced. Explicitly, the aim to alleviate the

difficulties being faced by coral reef researchers in data and information management and through the automation of processes in the data analysis functions, infer new knowledge. The Semantic Reef architecture combines Semantic Web and workflow technologies in an unconventional approach to find synergies for data manipulation and analysis.

Other eco-informatics initiatives were compared to position the Semantic Reef project. The analysis of the differing architectures was provided for each related project and, in contrast, the Semantic Reef architecture is shown to have a distinctive mixture and approach.

The following chapter describes in detail the methodology and development of the Semantic Reef Knowledge Base (KB) and the ontologies within.

Chapter Three

Developing the Ontologies

3.1. Chapter Synopsis

Ontologies are used in computer science to explicitly describe a concept so it is "computer-understandable" ergo "computer-processable" (Lassila and McGuinness 2001). Specifically, ontologies are well defined descriptions, with added contextual information, presented in a format a computer can use to make "intelligent" decisions. The contextual information adds computer processable "meaning" to the data which, in turn, can be employed by the computer to dynamically infer connections, both obvious and/or latent, between digital entities. There are a range of different types of ontologies from informal or "lightweight" through to formal or "heavyweight" and they are categorised based on the expressive complexity required to define the concept (refer to Chapter 2 §2.4.1.3). Accordingly, the type of ontology, or ontologies, to choose when designing a KR system depends on the desired outcome and the purpose of the system (Gomez-Perez et al. 2004).

The Semantic Reef KB consists of a hierarchy of ontologies that describe coral reef ecosystems. This chapter explains and justifies the methodologies involved in the design and development of the coral reef ontologies. First, a model of the functions of a generic coral reef is described from the perspective of a marine expert. Then, the expert's model is engaged as semantic "building blocks" to express coral reef concepts as ontologies within the KB development. The "building blocks" are used in the ontology development methodology to construct the KB at three levels: the composite, component and holistic levels, in order to maximise reusability and usability. Initially, ontologies are created to define simple ecosystem concepts at the composite level. These composite concepts interact with each other within a specific component, which is the next ontology layer. Then, explicit relationships are defined that link the components together at the holistic level. Together the ontological levels form a reusable hierarchical KB that describes any coral reef in the world regardless of location, makeup or type. A discussion of the design decisions and justifications of the hierarchical set of reusable and usable ontologies concludes the chapter.

3.2. The Coral Reef – a Domain Expert's Perspective

This thesis crosses the computer science and marine science disciplines and thus domain expertise was required for both. A domain expert or Subject Matter Expert (SME) is a person with specialist knowledge or skills in a particular area or discipline. If a computer is to understand an area of knowledge such as marine science as an ontology, it has to have certain knowledge about the languages, terms and concepts of that specialist area. Commonly, the computer expert will not have the vocabulary or expertise in the different field of knowledge in order to describe the concepts of that field as ontologies. Instead, a domain expert is required, one who sees and understands the specific field of study in a particular way and adopts a particular language to express that perspective. Accordingly, collaborations are necessary in cross-discipline applications, such as the eco-informatics project described here, to combine the perspectives of specialist disciplines with human/computer translation functionality.

The semantic translation of the marine expert's perspective of a coral reef ecosystem was essential to the development of the Semantic Reef KB. Coral reefs can be modelled from different perspectives; for example, as a functional model or as a community composite model, depending on the requirements of a specific study. Holmes (2008a), the marine domain expert here, provided a functional model of a generic coral reef (Figure 3.1). This model describes the basic functionality of a coral reef system and includes components such as coral reef community composition, nutrient dynamics and environmental and anthropogenic influences (Holmes 2008b). In fact, the diagram

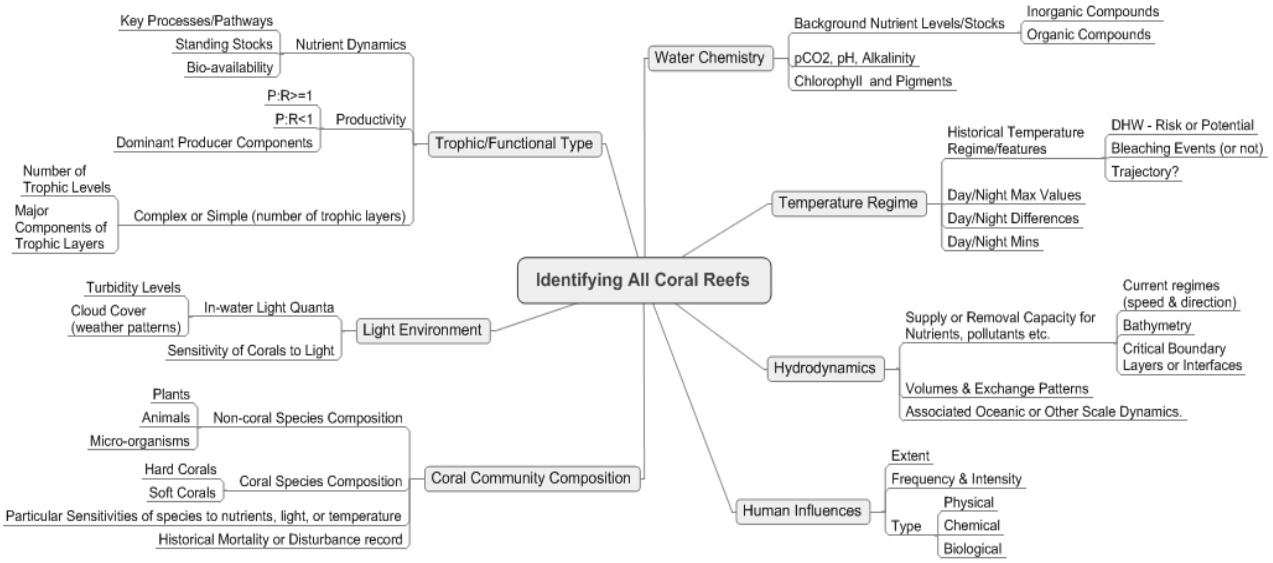


Figure 3.1 – Coral Reef functional concepts supplied from a marine expert – Each function has a natural hierarchy of sub-functions or related factors.

shows at a broad level the functions entailed in any coral reef independent of the reef type, where it is situated globally and what is contained in the community "mix".

Holmes's functional model is a hierarchy of concepts and is based on a holistic view of any coral reef (Davey et al. 2008). The hierarchy of functions begin with the main components, which are the first functional nodes of the model in Figure 3.1: the coral community composition, the trophic functions, the hydrodynamics, the water, chemistry, temperature regimes and the light environment. Each of these main components contains a hierarchical composite of features and conceptual terms. For instance, the light environment component contains two sub-nodes: sensitivity to light and an in-water light quanta factor. In turn, the in-water light quanta factor consists of two lower sub-nodes, which are the turbidity and cloud cover sub-factors (Figure 3.1). Therefore, each principal component of the model is a concept that can be defined independently as a composite of its sub-nodes.

The definitions of the principal components and their composite are transferable to the computer in a hierarchical ontological structure because they are based on relationships. The relationships can be defined at each level of the functional model, from the major components down to the smallest nodes, and include intra-relationships[23] and inter-relationships[24]. The intra-relationships are the relationships that are contained *within* each main component. In contrast, the inter-relationships are the relationships that span *between* the main components as they relate to each other at the holistic level.

This complexity of the relationships within a coral reef system makes it hard to adapt to a computer. Holmes's model does not concentrate on singular relationships of community composition or environmental aspects alone, but includes and segments all other functional factors that are prevalent to any coral reef; for instance, hydrodynamics, anthropogenic influences and trophic functions. To describe a whole reef system modelled in a single ontology would be extremely complicated; the intra- and inter-relationships descriptions would be too complex for it to be maintainable or reusable. A computer needs to understand the relationship between concepts to reason and infer new knowledge and can best do so in modules. Thus, if the coral reef entity is defined in modular fashion, the relationships will be less complex and easier to maintain. Ontology design methodologies support modularity (Rector 2003; Grau, Horrocks, Kazakov et al. 2008). The expert's model separates components in a hierarchical form and these components can be used as semantic "building blocks" for translation into a modular ontological form in the design of the KB.

[23] The prefix intra is used to determine the relationships within an entity.
[24] The prefix inter is used to determine the relationships between entities.

Consequently, a hierarchical ontology structure, from an atomic level to a holistic level, was created to represent the concepts, functions and distinct layered formation of Holmes's model. The atomic level focuses on the composite of factors that are part of each individual main component and the intra-relationships *within*. Conversely, the holistic ontology level focuses on the inter-relationships *between* main components themselves. The intra and inter connecting relationships were conveyable to the constructs used in the ontology languages to define concepts and functions.

At the atomic level, a standalone component contains a composite of internal concepts and their terms, factors and properties which can be described separately. For instance, the "human influence" component can be defined by the following composite of factors and properties: the type of influence, the degree of intensity, the frequency and the extent of the influence (refer to Figure 3.1). These factors can all be categorically described from low through to high and have interactive relationships with the other factors within the ontology. Explicitly, the impact factor of a "human influence" component is directly relative to the type of human influence, how intense it was, how frequently it occurred and to what extent it occurred. To illustrate, an oil spill, which is a "chemical" influence type, has quantifiable descriptions, such as intensity and extent levels (i.e., low, medium and high) which are categorically defined in the "human influence" ontology. The quantifiable descriptions and relational connections can be classified by a reasoning engine to automatically determine a categorical impact factor. Consequently, if the impact is high, then the system can also logically assume the damage on the reef to be high. The terms, factors and properties of one component also interact with the factors of other components at a higher holistic level (an oil spill may change light quanta or turbidity levels).

The holistic level expresses the external relationships that interlink the main elements of the model. The holistic viewpoint determines the many interconnecting, interwoven relationships in an ecosystem as a whole, in contrast to the atomic level, which identifies the composite of factors and internal relationships within the individual components. The inter-relationships that interconnect the main components of the expert's model entail cause and effect dynamics. For instance, there are consequences and connections between the "human influence" component and the other principal components in the model. In the case of an oil spill, the information introduced to the "human influence" component (i.e., impact factor, intensity and extent) changes properties within the "light environment" component, particularly the in-water light quanta level, as they are interconnected. Specifically, the oil spill's extent as a "chemical human influence" infers a change to the available light quanta within the "light environment" component.

The KB design required flexibility so the ontologies could be reused in different hypotheses. Many of the linkages and connections between entities in a coral reef domain, and within the ecosystem as a whole, are well researched. However, many factors still remain unstudied and often new results lead to defeasibility; that is, when what was once known as true is proven otherwise (IPCC 2007; Antoniou et al. 2001). Therefore, the relationship definitions between concepts in the ontologies must be flexible for future modifications, which may be required due to the introduction of new domain-specific information or knowledge. Also, modifications to the KB may be necessary in light of a new hypothesis. A modular ontology design was implemented so the intra- and inter-relationships, at both the atomic level and the holistic level, could be modified to suit the line of enquiry.

3.3. The Hybrid Ontology Design Methodology

There are numerous published ontology design methodologies available to assist in the representation of knowledge. Fernandez (2002), Corcho (2003) and Hadzic (2009) described the methods and techniques and identified the commonalties and differences of the most common current ontology engineering methodologies. However, to date there is no one preferred standard methodology (Corcho et al. 2003; Garcia et al. 2009). The choice of methodology or combination of methodologies depends on the purpose of the knowledge base (Gomez-Perez et al. 2004; Corcho et al. 2003). The main criterion for the methodology in this thesis was the degree of simplicity and flexibility to easily merge and/or separate the KB elements. The ability to separate elements of the KB, in particular the axiom definitions, instance data and rules, was required to ensure modularity and ultimately the reuse of the KB.

The strategy for the Semantic Reef ontological architecture was to maximise reusability and usability. To maximise reusability, the KB was required to be generic, modular and flexible and to remain independent of the hypotheses being posed. Alternatively, to be usable, the KB would require domain-specific data and information that is applicable to the actual reef and hypothesis in question.

Hence, a hybrid of design methodologies was adopted in the creation of the knowledge base. The hybrid model is based on intra- and inter-ontological development strategies. The intra-ontological design focuses on the concepts and relationships at the atomic level of the expert's model, so each main component is separately defined in ontological form. At the inter–ontological level the concepts and relationships that link all main components together, at the holistic level, are defined. The design methodologies used in this hybrid model include the *seven step knowledge*

engineering methodology (Noy and McGuinness 2001), Uschold and King's (1995) three strategies to indentify concepts and the *Developing Ontology-Grounded Methods and Applications* (DOGMA) approach (Jarrar and Meersman 2008). The first and second are a generic set of guidelines to construct ontologies of any concept and were employed for the intra-ontological development. The third offers a strategy that focuses on ontology reusability verses ontology usability and was adopted for the inter-ontological development process employed here.

3.3.1. The Intra-Ontology Development Methodology

The intra-ontology development utilised two distinct methods for the individual ontologies, one for procedural guidelines and the other for the class structure and definitions. The ontologies, based on the functional coral reef model in Figure 3.1 were designed separately and in collaboration with SMEs. The procedural steps of the methodologies were applied to guide the ontology design and collaborations during the development stages. Because each concept of the model has varying degrees of complexity and depth, the approach taken in the development of the class structures was important. For example, the light environment node of the model has a shallow class structure compared to the taxonomy of plants and animals, because of the range of terms involved in capturing the concept.

Noy and McGuinness's (2001) knowledge engineering methodology was adopted as a procedural guide. The seven step ontology development guideline is a methodology for creating ontologies based on declarative knowledge representation systems. The systematic steps begin with a declaration of the scope, with questions about the purpose of the ontologies and the domain they are to cover. Next, the domain vocabulary is defined for the individual ontologies by listing the relevant terms. The vocabulary is extended by defining the class and subclass structures and is followed by explicitly detailed relationships between the classes, properties and property restrictions. Finally, when the instances of the defined concepts are identified and included, the KB is created.

The hierarchical class structure development approach for each ontology was determined by the three methods proposed by Uschold and King (1995). The three strategies for identifying the main concepts in an ontology are: the top-down approach, which identifies concepts from the most generalised to the most specific; the bottom-up approach, which is the opposite (i.e., the most concrete to the most abstract); and the middle-out approach, where the most relevant concept is identified first and then generalised and specialised (Uschold and King 1995; Uschold and Gruninger 1996).

Factors such as the over complication or simplification of the definitions and class structures assist in the choice of which approach is appropriate to implement (top-down, bottom-up or middle-out). A bottom-up approach requires a very high level of detailed terms defined first, which increases the overall effort due to the inclusion of unnecessary terms. The declaration of the most general terms first in the top-down approach results in better control of the level of detail but can lead to important factors being overlooked, which may result in the incomplete definition of a concept. Further, both top-down and bottom-up make it difficult to spot commonality between related concepts, which increases the risk of inconsistencies and can lead to counter-productivity and more effort. Because the middle-out approach maintains a balance in the level of detail it is less prone to these problems. The detail is specialised or generalised only as required when defining a concept in the middle-out approach, which leads to less re-work and a higher degree of stability in the design (Uschold and King 1995; Uschold and Gruninger 1996; Uschold 1996). The complexity of the reef ecosystem concept, and the degree of domain knowledge available to capture the concept or declare the class structure decided the specific approach chosen to model the separate ontologies.

Consequently, the class hierarchies of the individual ontologies were developed using both top-down and middle-out approaches. The top-down approach is appropriate for concepts that have a distinct set hierarchy of knowledge such as the scientific categorisations of species. Hence, the top-down approach was taken when developing the coral reef plant and animal taxonomy, because cataloguing from kingdom to species in both scientific and common names affords a natural rising. Alternatively, the middle-out approach is appropriate in circumstances where the scope of the concept being described is unclear. Hence, all other ontologies were designed using the middle-out approach, for instance, the "Human Influence" ontology, where the biological, chemical or physical influence could be declared and systematically specialised (type, frequency and intensity) and generalised (deliberate or accidental).

3.3.2. The Inter-Ontology Development Methodology

The inter-ontology development adopted a DOGMA strategic approach. DOGMA is a methodological ontology engineering framework that divides an ontology into two specific axiomatization concepts: the domain and application axiomatizations (Jarrar and Meersman 2008). The domain axiomatization characterises the meaning of the domain vocabulary (for reuse) and, in contrast, the application axiomatization determines how the vocabularies are applied/adopted.

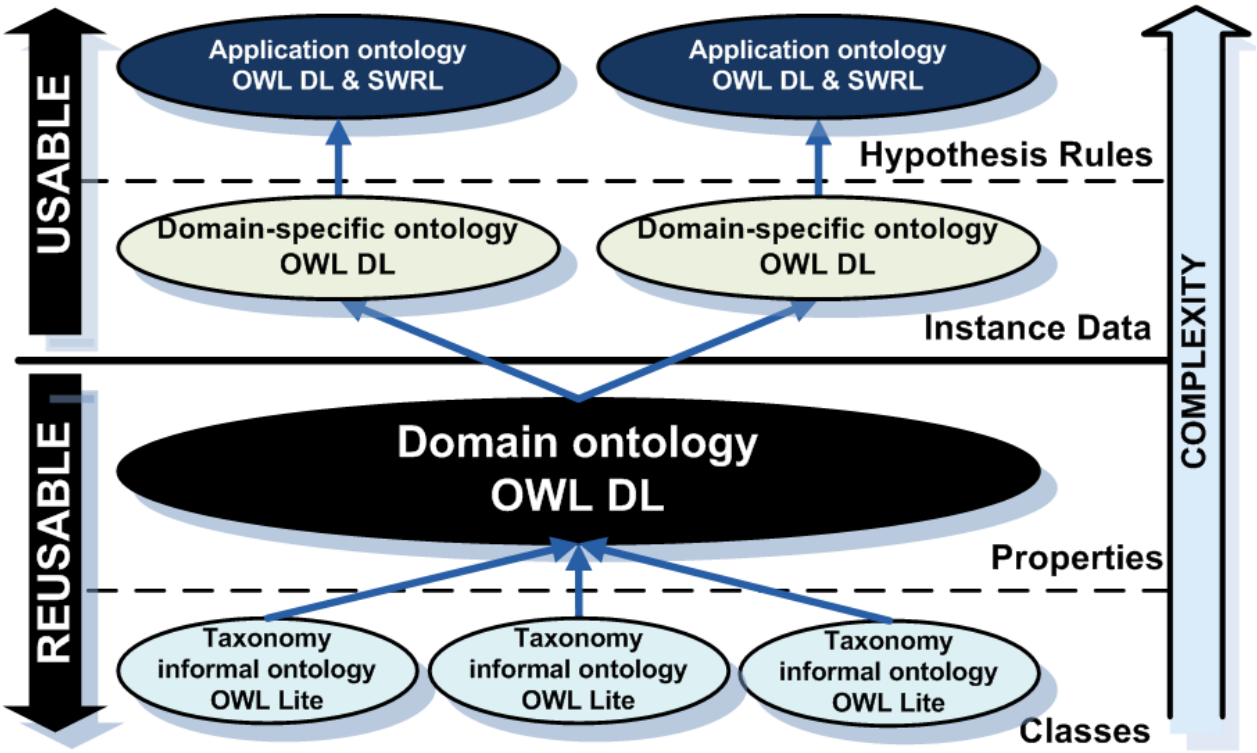


Figure 3.2 – The inter-ontology methodology supports simultaneous reusability and usability by separating the domain ontologies from the applications ontologies.

Consequently, the more emphasis on holistic domain perspectives, the higher the reusability and the more emphasis placed on application requirements, the higher the usability.

According to the DOGMA approach, the separation of the domain knowledge from the application or domain tasks enables both reuse and usability. This DOGMA reusability philosophy was suitable for the Semantic Reef KB for hypothesis-driven research and was adaptable to the ontological design. The ontologies were designed as "reusable" domain ontologies, to describe coral reefs generically, and as "usable" domain-specific and application ontologies, to describe specific coral reefs and the rules of the hypotheses about that reef. The ontologies are imported from the lowest reusable layers upwards with two distinct levels: the reusable domain ontologies and the usable domain-task and application ontologies (Figure 3.2).

The implementation of the DOGMA philosophy effectively separated the domain (reusable) ontologies (i.e., the "Coral Reef" ontology) from the higher application (usable) domain-task ontologies (i.e., the specific reef hypotheses) (Figure 3.2). The reusable components of the ontology base are contained within the "Coral Reef" generic ontology. This ontology imports all lower ontologies and maintains OWL DL axioms for richer relational descriptions that are indicative of any coral reef. The usable component of the KB lies in the domain-task ontologies,

which import the "Coral Reef" ontology and are populated with instance data pertinent to the specific reef system and hypothesis (e.g., GBR.owl, Moorea.owl, Bahamas.owl, etc.). For example, there are distinct levels of common and uncommon features in questions of data from different reefs. The sensor data from the GBROOS Davies Reef weather station on the GBR or the Coral Reef Environmental Observatory Network (CREON) group conducting research in the French Polynesians may have common aspects such as environmental and geographic measures, but also unique components such as regional phenomena or instrumental measurement device specialities. The elements and classes of the generic "Coral Reef" domain ontology are the same for either location. However, at the usable level of the hierarchy, the instance data differs at the domain-specific level (i.e., the "GBR" or the "Moorea Island" ontologies) and rules of the application ontologies differ due to the actual questions asked of the system (i.e., "GBR Rules" and "Moorea Island Rules" ontologies).

3.4. Describing Coral Reefs as Reusable and Usable Ontologies

Concepts can be modelled through ontologies in a variety of ways and in varying degrees of granularity to serve a distinct purpose (see Chapter 2 §2.4.1.3). In ontological design, the scope of complexity spans a range of data models. The data models start with vocabularies or a thesaurus of a domain concept, and extend to the complex formal ontologies that incorporate logic systems for inferring new knowledge autonomically. The choice of ontology is determined by the extensibility and expressiveness required; that is, by the information or knowledge it is designed to produce (Gomez-Perez et al. 2004; van Heijst et al. 1997; Chandrasekaran et al. 1999).

The more complex the ontology, the more prone it is to inconsistencies and the higher the restriction on flexibility and reusability, independent of applications. The simpler constructs of the informal "light-weight" ontologies, such as taxonomies and common ontologies (described in §2.4.1.3.1), foster the highest degree of flexibility in the ontology design. As more expressive descriptions and constraints on the relationships are required to define complex concepts, a trade-off between flexibility and semantically rich definitions occurs. Here the ontologies move to more formal constructs as logic systems such as DL, are introduced (Figure 3.3) (Gomez-Perez et al. 2004).

The types of ontologies chosen, and the level of formality, were based on usage requirements and the depth needed to define each concept. The distinction between the reusable

domain ontologies and the usable application ontologies was determined by the level of granularity required to describe the concept or achieve the purpose. The purpose was established by the degree of use and reuse required. The distinction of ontologies for use and reuse lead to the effective separation of the classes and properties, which are contained in the domain ontologies, from the instance data (domain-specific ontologies) and rules (application ontologies). The classes and properties make up the domain ontology base; hence, each component of the domain expert's model (Figure 3.1) was defined as a single or group of ontologies. The level of formability for each was decided by the complexity of the relationships required. Firstly, the relationships between the entities of each separate component of the model, at the atomic level were explicitly defined and then, at the holistic level, the relationships between the components themselves.

A "ground-up" physical architecture is the resulting hierarchical formation of the KB (Figure 3.3), not to be confused with Uschold and King's (1995) "bottom-up" strategy for design. The "ground-up" physical flow consists of the base-level light-weight ontologies imported to the more complex DL ontologies. Then the task-specific and application ontologies, at the higher

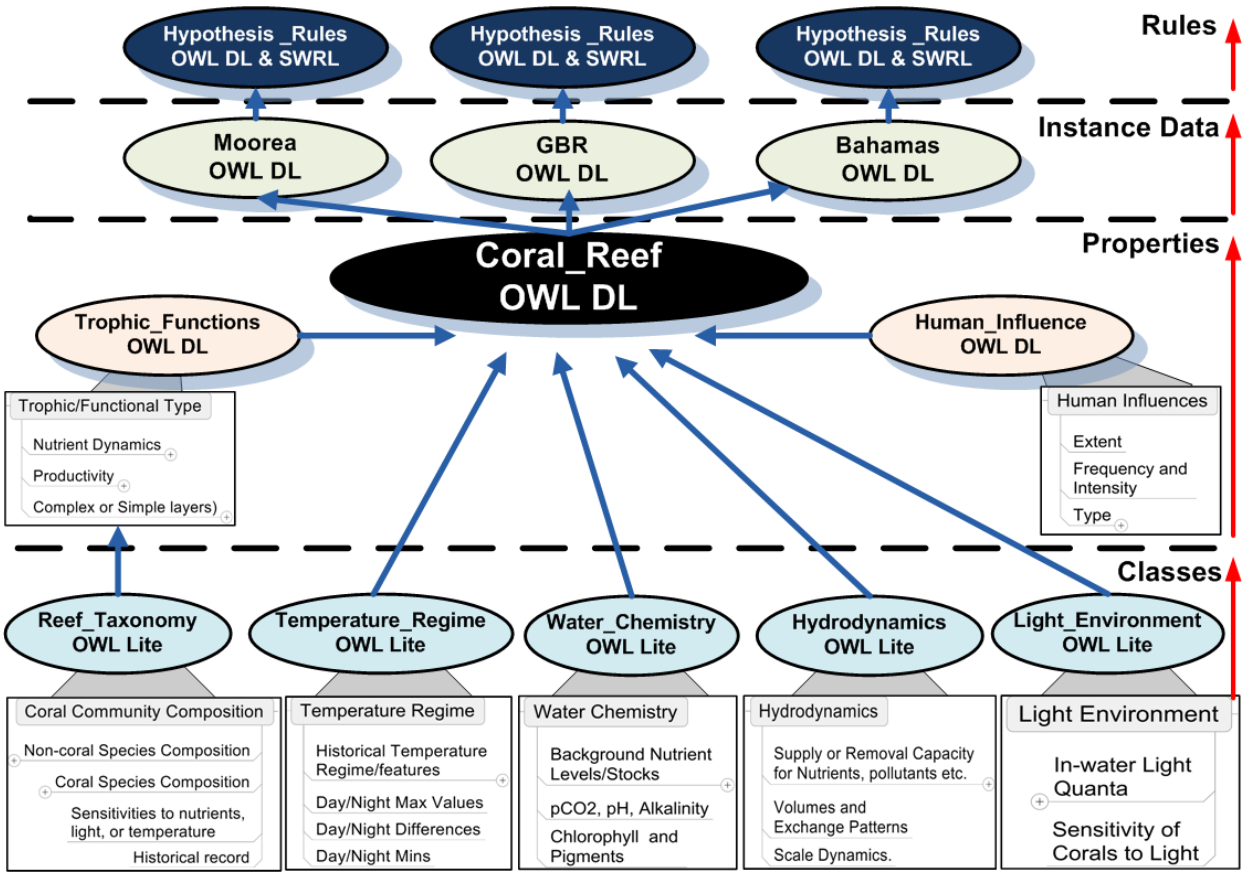


Figure 3.3 - Coral Reef concepts segmented into a hierarchy of informal to formal ontologies.

levels, employ the reusable coral reef ontology base beneath. Finally, at the highest level, finely detailed inference rules can be written to the system as propositions. The rules infer conclusions about a specific problem on a particular reef, regardless of location (e.g., a coral bleaching watch on the GBR or coral spawning events on the Moorea Island reefs).

3.4.1. The Base Level –Define the Coral Reef Domain Vocabulary

The first level of the ontology constructs are the lightweight domain vocabularies, glossaries of terms and thesauri. To construct the KR system the first step was to establish the terms of the domain in the form of a controlled vocabulary. A controlled vocabulary is a set of restricted words or terminology used internally, by an information community, to describe resources or discover data (i.e., the jargon) (Garshol 2004; Neiswender 2009).

Often there are several terms to describe a single concept (i.e., repetition). Interoperability between the many different data repositories is difficult because keyword queries generally work by literal string comparisons searching highly structured disparate data schemas. Hence, a complete list of redundant terms is required to associate all synonyms for one single concept.

A more complete controlled vocabulary of a coral reef requires inclusion of both common and scientific names. For example, the terminology to name coral entails common terms such as soft and hard coral, as well as scientific terms such as the *Anthozoa* sub-classes: *Alcyonaria* (soft corals, sea pens) and *Zoantharia* (sea anemones, stony corals, black corals), which are also referred to in some databases as *Octocorallia* and *Hexacorallia*, respectively. If the disparate data sets are to be bridged, all terms within a reef community composition must be listed, including both common and scientific names. The class structure of the ontologies commences after all possible terms are listed.

Organisations exist whose primary focus is to standardise the data and metadata format produced in marine research. A prime example is the Marine Metadata Interoperability (MMI) project, which is an NSF funded initiative that aims to "promote the exchange, integration and use of marine data through enhanced data publishing, discovery, documentation and accessibility" (MMI 2009). The MMI lays a foundation, via a comprehensive set of guides, for scientists and data managers to create metadata (Stocks et al. 2009). These guidelines include advice on good metadata practices, forums for collaboration and web applications for working across marine data systems. MMI have extensive lists of metadata repositories, controlled vocabularies and current standards references (e.g., ISO 19115 *Geographic information – Metadata*). Also, a list of marine

related ontologies is made available for projects such as the Semantic Reef to access, import and, if need be, modify for specific purposes.

A multitude of sources are available to form a controlled vocabulary of a marine ecosystem. The MMI site hosts a list of controlled vocabularies of biological taxonomies. For example, the Interagency Taxonomic Information System (ITIS)[25], Project 2000[26] and the Global Biodiversity Information Facility (GBIF)[27] are collaborative efforts to standardise the biological taxonomy information, with a comprehensive database offered by each. Importantly, there are e-Science endeavours to defragment the biological taxonomies currently dispersed in both online and ink-on-paper publications (Page 2006; Clark et al. 2009). For example, the Universal Biological Indexer and Organiser (uBio)[28] is an initiative within the science library community to join international efforts (i.e., ITIS, Project 2000 and GBIF) to create a comprehensive catalogue of known names of living and extinct organisms.

A controlled vocabulary for the coral reef domain includes a vast number of plant and animal species. Currently there are 11,106,374 biological names indexed by uBio, over 1,140,000 in the *Animalia* Kingdom alone. Also, there are 31,100 species and 276,100 common names logged in FishBase[29], a global information system about fish. Due to the extensiveness of flora and fauna, a sample of the controlled vocabulary for a coral reef domain was developed for the implementation of the Semantic Reef project, as a proof-of-concept. The sample forms the base level community composition ontology, which is a non-comprehensive list of species with both scientific and common names, for a coral reef environment. The list can be expanded in future additions to the KB by simply extending the "Reef" taxonomy class structure.

3.4.2. The Base level Ontology Language - OWL Lite

The ontologies at the lowest level of the hierarchy are designed for maximum flexibility and reusability. Automated classification and inference across disparate data is also important at this level; therefore it is apt the system be maintained as a DL KR system. OWL DL and OWL Lite, which is a subset of OWL DL, are both based on DL so all inferences available in an OWL DL or OWL-Lite ontology can be computed by the reasoning engine. The ontologies at the base level are simple concepts and were created with OWL-Lite, which is the lowest subset of OWL Full

[25] http://www.itis.gov/
[26] http://www.sp2000.org/
[27] http://www.gbif.org/
[28] http://www.ubio.org/
[29] http://www.fishbase.org/

but is expressive enough to define the simple taxonomies, or the hierarchies of classes, that were required at this level. Also, because only simple constructs are available in OWL Lite it does not use as many computing resources when the reasoning engines are applied.

OWL Lite provides the following basic ontological constructs for simple decidable ontologies (Lacy 2005):

- OWL classes can be derived from other OWL classes (i.e., subsumption) with the sub-class constructs rdfs:subClassOf and owl:equivalentClass;
- Individuals can be declared equivalent or different with the owl:sameAs, owl:differentFrom and owl:AllDifferent constructs (i.e., synonymous and antonymous relations);
- Property characteristics can be declared with owl:FunctionalProperty, owl:InverseFunctionalProperty, owl:TransitiveProperty and owl:SymmetricProperty, to define inverse (binary relationships), transitive (a property of a superclass must be a property of a subclass), symmetric (links individuals from a domain and a range inversely) and functional (one only property) relations;
- Restrictions on properties can be declared with existential and universal quantifiers (i.e., owl:someValuesFrom and owl:allValuesFrom) to constrain *some* or *all* of the property values and individuals that can belong to a particular class;
- Properties can be manipulated to constrain limited cardinality of an individual (i.e., owl:cardinality, owl:minCardinality or owl:maxCardinalty all limited to 0 or 1); and
- Classes can be described as the intersection of another class using the set operator owl:intersectionOf, which can be viewed as the logical conjunction "AND" (Smith et al. 2004).

3.4.3. Base Level – The Informal Taxonomies and Lightweight Ontologies

The base level taxonomies and general ontologies are developed in this stage by adding structure to the controlled domain vocabularies. In this development stage, the controlled vocabularies specified in the first step of the ontological design are converted to form a base-level generic community composite taxonomy and the base-level general environmental ontologies.. In effect the taxonomy is simply an ontology where only the class structure is declared without

restrictions and complex properties or relationships. That is, only equivalencies (i.e., synonymous relations), diversities (i.e., antonymous relations) and classifications of generalisation and specialisation (i.e., hyponomous or "is a" relations) are declared here.

The less complex constructs employed in a taxonomy support flexibility in design and reuse. For example, the community composition of animal and plant species does not require highly complex relational properties, only properties to define "same as", "different from" and "is a" relationships for subsumption and classification of instance data. Therefore, a taxonomy is the most appropriate semantic structure because the simplicity facilitates ease of reuse for any hypothetical question asked of the system. Further, the simple constructs of the taxonomies use fewer computer resources by reasoning engine, which is significant in the scalability of the KB.

3.4.3.1. The Base Level Reef Community Taxonomy

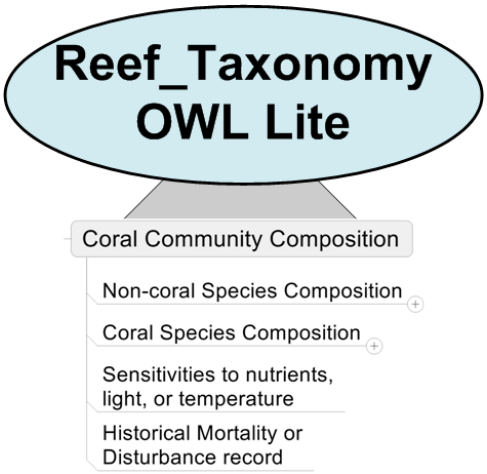


Figure 3.4 – Base level OWL Lite Reef community taxonomy.

The community composition "Reef" ontology was created as a base-level taxonomy (Figure 3.4). Written in OWL-Lite, the "Reef" Ontology lists the vocabulary of the coral reef domain in both scientific and common names. The only relationships required for this hierarchical classification level are the owl:equivalentClass and the rdfs:subClassOf, which explicitly state the equality or "is a" relationships between the phylum and family names and their common names. This taxonomy was designed with the "top-down" design approach, as discussed in section 3.3, because it has a natural rising of highly structured concepts. The hierarchy of relationships of the reef concepts were obtained from pre-existing detailed taxonomic resources. Also, future additions to the KB are simplified because the complexity of relationships is minimal. For instance, newly discovered species can be easily added or the taxonomy can be extended with extra information from the biological databases mentioned earlier.

Queries and inference over once disconnected data is now possible because the community composition taxonomy links between disparate datasets to populate the KB. For example, *Acropora* in one database might be staghorn coral in another and upon classification by the reasoning engine, the asserted instances from one database will belong to both scientific and common name classes automatically.

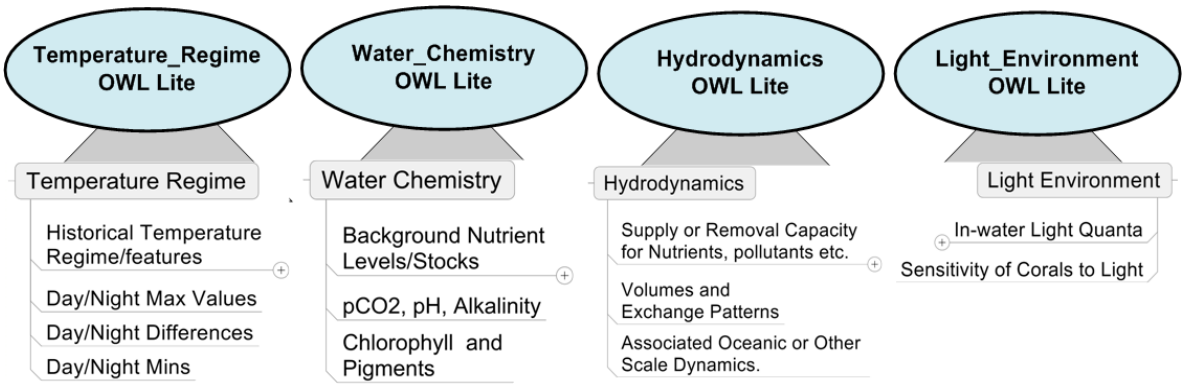


Figure 3.5 – Base level OWL Lite Environmental Ontologies.

3.4.3.2. The Base Level Environmental Domain Ontologies

The base-level environmental ontologies are informal domain ontologies to describe the environmental elements pertinent to coral reef systems. The four main quantifiable environment elements (the temperature regime, water chemistry, hydrodynamics and light environment components) are depicted in common lightweight informal ontologies (Figure 3.5). Only simple data-type properties are required to describe factors that are significant to these environmental regimes. Data-type values such as integers, floats, strings and Boolean constructs that can represent the quantifiable environmental values (e.g., SST, humidity percentile, pH level, etc.).

These lightweight ontologies, written in OWL Lite for simplicity, can be populated using the Kepler workflow with historic or real time data. For instance, SST from geospatial satellite data from NOAA or remotely sensed data from GBROOS. The base level environmental ontologies are then imported into the more complex domain "Coral Reef" ontology.

The reuse of existing ontologies was considered and applied where appropriate. However, the strategy was to structure the Semantic Reef KB on the functional coral reef model (Figure 3.1) supplied by the domain experts (Davey et al. 2008). The environmental domain ontologies were developed based on the expert's model as semantic "building blocks" with the middle-out methodological approach. To date, the temporal ontology developed at Stanford University for use in the SWRL rules has been implemented (O'Connor et al. 2006) and the future development of the ontology knowledge base will incorporate or map other common ontologies to the system.

Many pre-existing ontologies were not suitable due to class structure or complexity when applying them to the intra and inter-ontology framework of this KB. The choice of class versus instance depends on the purpose of that ontology. The KB consists of temporal instances of a reef that describe the community and environmental makeup at a particular moment in time and place.

To express the quantifiable values of the environmental elements (e.g., SST, pH, etc.) as data-type properties asserted to the reef instance, for a temporal moment, was an appropriate technique. The alternative would be to declare one reef instance (e.g., Davies Reef) and link it to a numeric instance of a temperature class via an object property assertion (e.g., owl:hasValue). This would require the temperature class to be filled with all possible numeric values. Pre-existing common ontologies express concepts as classes in the class structure or as instances of a class that did not align to the class structure here. For example, the common ontologies that describe quantifiable concepts, such as units of measurement, express the variety of units as individuals and not data-type properties (NASA 2009; SEEK 2009), which are not appropriate for use in this KB.

3.4.4. The Description Logic (DL) Level

The higher level reusable ontologies introduce complexity of relationship descriptions among the concepts. The richer semantic layers of DL and inference rules are required to describe the intricate inter-relationships between the entities of a coral reef ecosystem. Although lightweight ontologies are appropriate to define the lower levels of the coral reef concepts, they are not extensible enough to manage the finer, more complex relationships that exist within an ecosystem. More complex relationships require more restrictions on properties to describe them. DL constructs offer functions, such as existential and universal quantification, cardinality and Boolean combinations, to describe the intricate inter-relations and ramifications of cause and effect between concepts. The base level common ontologies are then imported to the higher level OWL DL domain ontologies and more complex descriptions are applied for reasoning and inference.

Defined classes have explicit axioms declared that assert a necessary and sufficient relationship and only defined classes are reasoned over by the reasoning engine. A class or individual must meet the specified necessary and sufficient restriction on the property to belong to an inferred class. Therefore, the choice of whether a concept should be declared an actual class (TBox), or an individual of a class (ABox) is determined by the desired classification and subsumption outcome as a result of the logic system.

3.4.5. The Higher Level Ontology Language - OWL DL

OWL DL extends OWL Lite with additional decidable language constructs (McGuinness and van Harmelen 2004). OWL Lite provides support for limited representation of information and because it has fewer available constructs, it is simpler to understand, implement and less taxing on

software and hardware requirements. However, finer granularity is possible with DL when more extensive descriptions of relationships and concepts are required.

OWL DL has all of the functionality of OWL Lite but with fewer constraint restrictions and more available constructs. The constraint relaxations include:

- The capability of full Boolean combinations; and
- The lifting of the restriction on the cardinality construct so the values are not limited to 0 and 1.

The additional constructs include:

- Extra class axioms such as enumerations (owl:oneOf), which define a group or list, and disjointedness (owl:disjointWith), to declare classes that cannot have the same individual as a member (e.g., gender classes are disjoint);
- All Boolean combinations of classes and restrictions are added (owl:intersectionOf, owl:unionOf, owl:complementOf), which can be viewed as representing the AND, OR and NOT operators; and
- The property restriction owl:hasValue, which adds filler information to restrict the value of a property linked directly to a specific individual.

OWL DL's existential quantifier owl:someValuesFrom (some) and universal quantifier owl:allValuesFrom (only) property restrictions close off possible ambiguities of a concept or a given property. To clarify the OWL meaning of these restrictions, owl:someValuesFrom equates to "at least one value of the property must be of a certain type but others might exist" whereas owl:allValuesFrom equates to "all values of the property must only be of a certain type or have null values" (Rector et al. 2004).

Reasoning engines apply axioms to infer connections in a DL KB. A DL KB consists of sets of axioms, which are statements of truisms or logical predicates that are used by the reasoner to make classification decisions. The reasoning engine automatically classifies, or "untangles", the KB to infer a class or an individual to belong to numerous classes dependent on the explicitly asserted axioms. The automatic classification and subsumption for this KB are handled by the reasoning engines: Pellet, RacerPRO and Fact++ (Mindswap 2007; RacerPRO 2008; FaCT++ 2008).

The more complex ecosystem concepts are part of the higher DL level ontologies and were created using the more expressive functionality of OWL DL. The level of granularity of ontologies

depends on its purpose and what information or knowledge the ontology is designed to produce (Gomez-Perez et al. 2004). Therefore, the complex concepts of Holmes's model such as the trophic layers and human influence components were created with the more expressive OWL DL.

3.4.6. The DL Level – Formal Domain Ontologies

3.4.6.1. The Trophic Functions Ontology

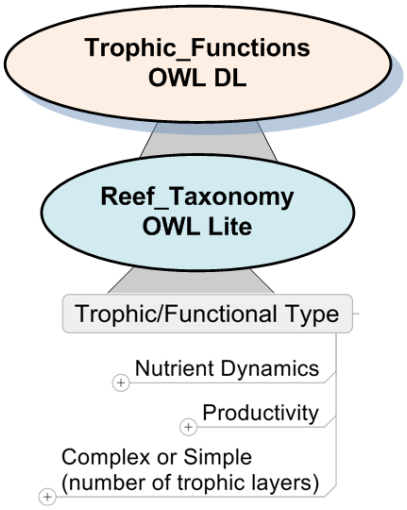


Figure 3.6 – OWL DL level Human Influence ontology.

Here, the "Reef" taxonomy is imported to the OWL-DL "Trophic Functions" ontology (Figure 3.6). At this higher DL level of the KB hierarchy the full relationships between the concepts (e.g., species) are expressed with the axiomisations and quantifications. Then, classification and subsumption of classes and/or instances is decidable autonomically by the reasoning engine. For example, the food web in which all species that are concurrently both carnivorous and herbivorous would be classified to the "omnivore" class.

Axioms that are common in OWL DL are class, disjoint and closure axioms. OWL classes overlap by default if not asserted otherwise, which is the OWA. Disjoint axioms are applied at the class level to explicitly state an individual cannot belong to another class simultaneously. For example, in the "Trophic Functions" ontology, the "Carnivore" class is declared disjoint to the "Herbivore" class because individuals from the "Carnivore" class cannot possibly belong to the "Herbivore" class. No disjoint axiom is asserted to the "Omnivore" class because it can contain individuals from either of the other classes simultaneously. Conversely, if there were no disjoint statement declared for the "Carnivore" and "Herbivore" classes, they would overlap and allow individuals to belong to both, which would lead to incorrect inferences by the reasoner.

Because OWL is an OWA environment, descriptions of classes should be "closed off" where appropriate; these are known as closure axioms. Closure axioms are a way of disambiguating a concept, leaving no opportunity for a wrong assumption. For example, one could describe the concept of an herbivore as "an animal that, among other things, eats some autotrophs". So the OWL statement:

```
Class Herbivores Defined ⊑
restriction(eats ∃ owl:someValuesFrom(Phytoplankton OR Algae))
```

will result in any class or individual that fits these conditions, will be classified a member of the "herbivore" class. However, unless explicitly stated, due to the OWA, any individual that eats plant life will be subsumed to belong to the "herbivore" class, including an omnivorous individual, which is not quite accurate (carnivores are also omnivores). Therefore, to remove the ambiguity, an explicit statement is required and the closing axiom is added to the previous OWL statement (refer to Figure 3.7) to further constrain the interpretation of the herbivore concept:

```
Class Herbivores Defined ⊑
Restriction(eats ∃ owl:someValuesFrom(Phytoplankton OR Algae))
           (eats ∀ owl:allValuesFrom(Phytoplankton OR Algae))
```

The reasoner would assume herbivores could eat animal life without the final closure axiom "owl:allValuesFrom" (i.e., only values from), because there were no statements to say otherwise. This is due to the OWA, which assumes if a fact is not there, it is unknown, not false (Rector et al. 2004). Once all disjoint and closure axioms are explicitly asserted the individuals added to the KB will be classified to the trophic layers dependent on the owl:objectProperty "eats" or its inverse property "is eaten by" (Figure 3.7).

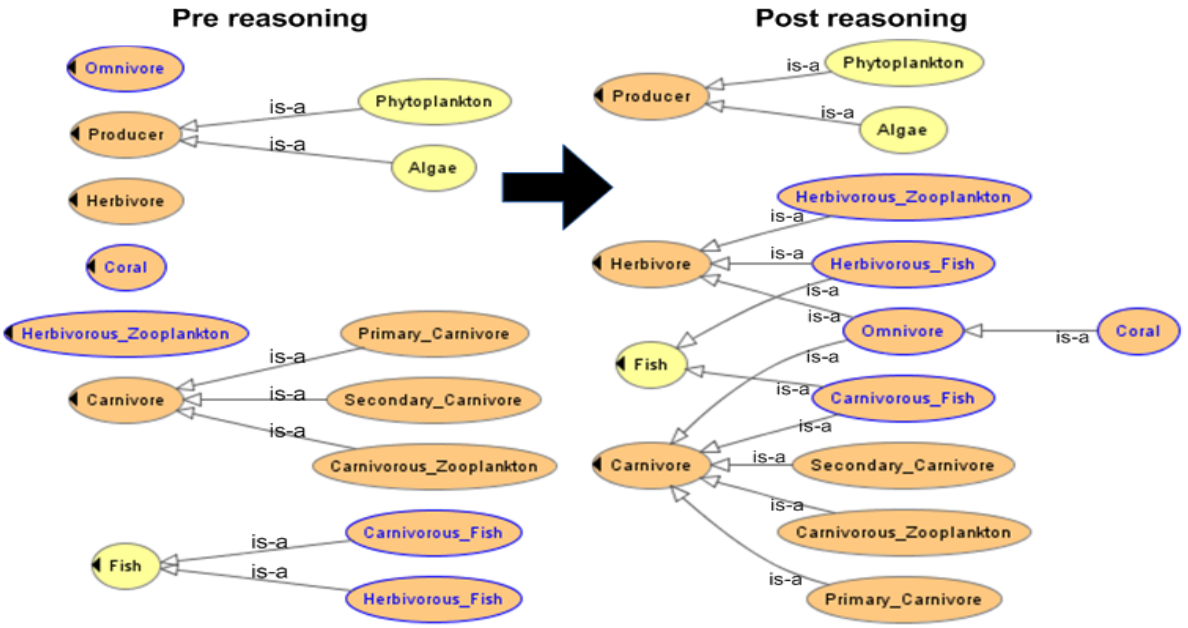


Figure 3.7 – The omnivore class after reasoning and subsuming

3.4.6.1. The Human Influence Ontology

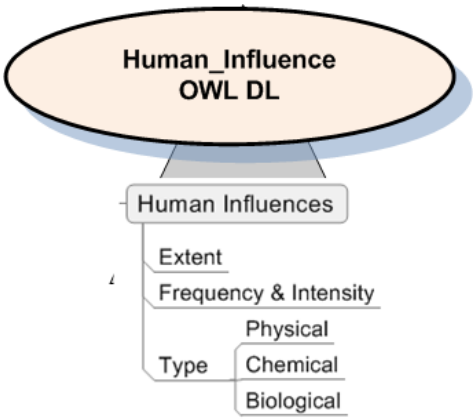


Figure 3.8 – OWL DL level Human Influence ontology.

The human influence component of the expert's model was also defined as an OWL DL ontology because of its complex intra-relationships. To describe human influences on coral reefs in a way understandable and decidable by a computer is not a trivial task. As discussed in section 3.3, the "Human Influence" ontology was engineered with a middle-out structural design approach that would enable future additions (Figure 3.8).

Three main categories of influence types (biological, chemical or physical) were declared and then defined by their specialisations and generalisations. The specialisations of the occurrence types were expressed by their levels of severity. The descriptive factors of each human influence type are the intensity, frequency and extent which, in turn, were expressed in ranges of minimum to maximum.

Human influence occurrences can be automatically subsumed to an impact level by the reasoning engine. Specifically, the "Intensity" and "Frequency" classes were created as enumerated classes with named individuals from low to high. Subsequently, the minimal to maximal affects were created as subclasses of "Influence_Extent" and class axioms of necessary and necessary and sufficient conditions were explicitly stated for automated reasoning. For example, to belong to the class "Chemical_Physical_Affect_Bad", a subclass of "Maximal_Affect", the conditions' state:

```
Class Chemical_Physical_Affect_Bad Defined ⊑
Restriction (hasHumanInfluence ∃
                owl:someValuesFrom(Physical_Influence)
            (hasHumanInfluence ∃ owl:someValuesFrom
                (Chemical_Influence))
            (hasHumanInfluence ∀ owl:allValuesFrom
                (Physical_Influence OR Chemical_Influence))
            (hasInfluenceType ∋ owl:hasValue (Physical))
            (hasInfluenceType ∋ owl:hasValue (Chemical))
            (hasFrequencyOf ∋ owl:hasValue (High_Frequency))
            (HasIntensityOf ∋ owl:hasValue (High_Intensity))
```

The KB is populated dependent on the pre-processing triggers from the Kepler workflow. For example, a highly frequent value can be determined as the data is presented in real-time and

properties such as "hasIntensityOf" and "hasFrequencyOf" are filled when populating the "Influence" class with instances. Then, all influence instances (such as an oil spill, dredging, pollution, flumes, etc.) will be automatically categorised to the severity of the affect following classification by the reasoning engine.

3.4.7. The Domain Ontology Level – The Reusable KB

The parent "Coral Reef" domain ontology is the highest level in the hierarchy that is reusable (Figure 3.9). The KB can be reused for different hypotheses. Each new hypothesis can be unique simply by importing the "Coral Reef" ontology to an application level ontology and then populating the KB at that level with relevant instance data. The "Coral Reef" ontology imports the lower environmental ontologies and "Reef" taxonomy, via the "Trophic Functions" ontology, and the higher level ontologies ("Human Influence" and "Trophic Functions"). All inter-relationships and connections between each of the reef concepts, which are the separate components of Holmes's model (Figure 3.1), are explicitly defined to form the holistic view of a generic coral reef. The inter-relationships between the ontologies, at this holistic level, are defined through axioms and restrictions on the properties of the "Coral Reef" ontology. The highest level of granularity is reached without sacrificing the capacity for reusability by adding the relationships that interlink the ontology's classes at this level.

This level of the KB is designed to maximise both reusability and also flexibility. The definitions and descriptions at this level of the hierarchy are generic and indicative of all coral reef systems at any global coordinates. However, if changes to the ontologies are required, the semantic technologies support modifications to the KB. Semantic technologies allow for changes to be easily incorporated into the KR system. Changes that are non-trivial in a relational database are trivial in a Semantic KB, and thus changing the relationships or adding new relationships to the schema is a simple task (refer to Chapter 2 §2.4.1.5). If an ontology within the KB hierarchy

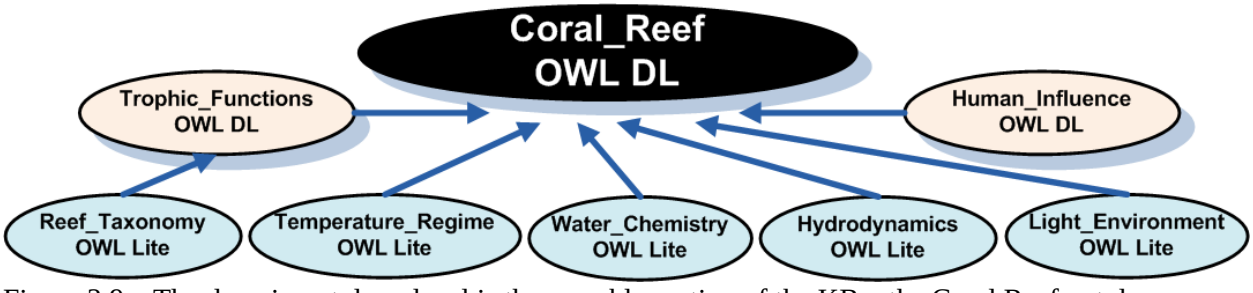


Figure 3.9 – The domain ontology level is the reusable section of the KB – the Coral Reef ontology.

required changing due to new information or new beliefs in the field of study, the property restrictions relevant to the relationships affected could be easily modified. For instance, if a new discovery found a compound that accelerated coral growth, and was previously unknown, it would change the trophic layers and growth rate relationships of the specific coral, which in turn would change the recovery rate of a bleached reef. The required modifications would be new property assertions and axioms to describe the changed relationships. The instance data would not be compromised because it is separate from the property assertions. More explicitly, the instance data is held in the domain-specific ontology at the higher application level, which is separate from the lower domain ontology that contains the property assertions. Reuse and flexible design is maximised because the logical domain descriptions are effectively separated from the instance data.

3.4.8. The Domain Specific Level – The Usable KB – The Instance Data

The semantic reef model is designed as a tool for observational hypotheses or casting alerts of any coral reef in the world. Coral reefs world-wide have an area mass of 284 300 sq km according to the most recent figures (Spalding et al. 2001) and the current monitoring and data collection efforts from government and private organisations are diverse and non-standard (Figure 3.10). The capacity for reuse relies on the ability to introduce new or different data into the system dependent on the hypothesis and the reef in question. Each new hypothesis may be a different line of enquiry and require different data from a preceding one. The KB can be refilled depending on

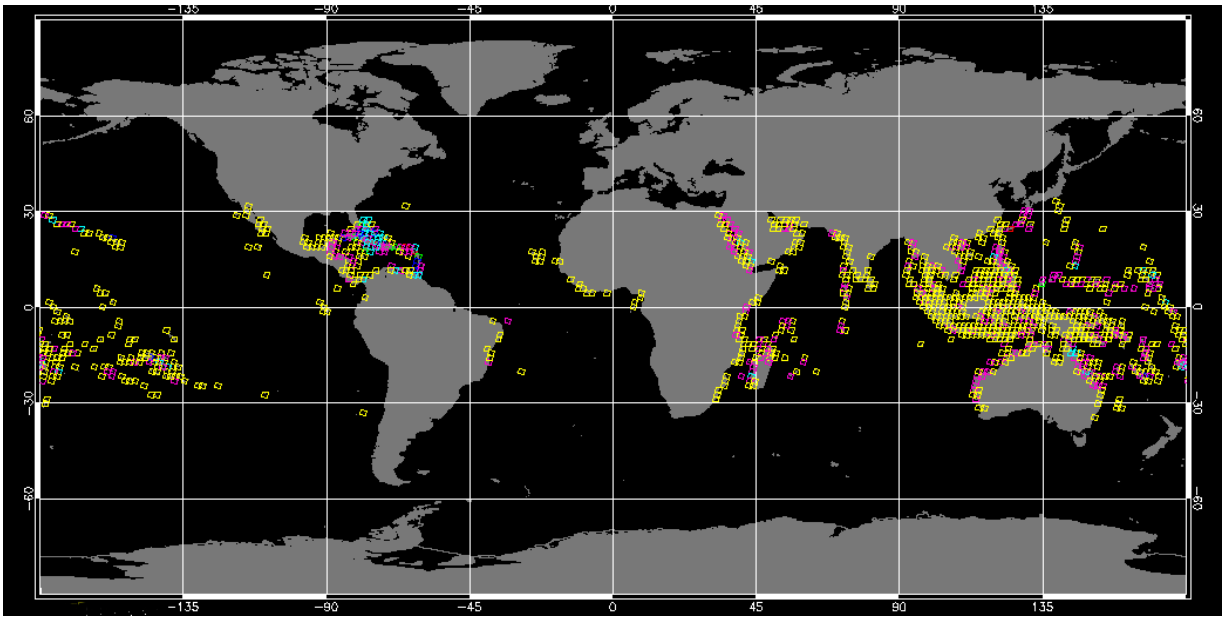


Figure 3.10 – World Map of Coral Reef locations correlated by the Institute for Marine Remote Sensing, University of South Florida (IMaRS 2009; Spalding et al. 2001)

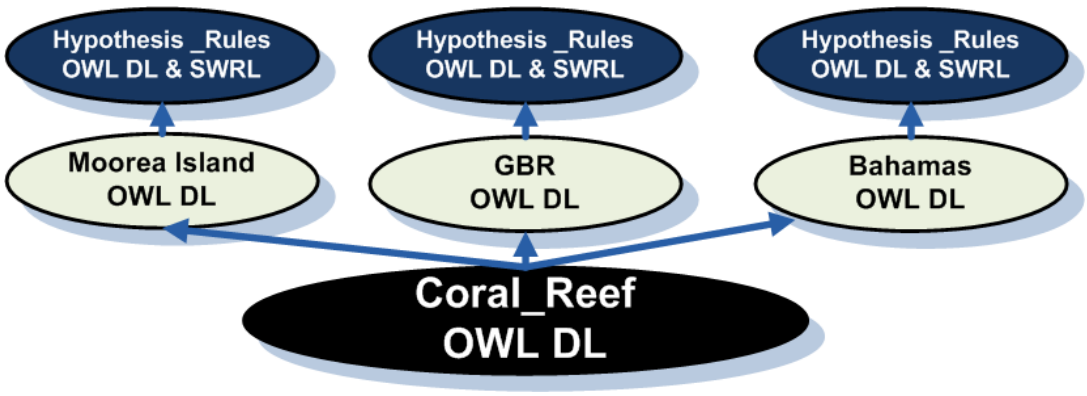


Figure 3.11 – The application ontology level is the usable section of the KB – domain-specific reef ontologies and rules ontologies.

the line of enquiry by separating the "reusable" domain KR from the "usable" instance data, which is a technique of the DOGMA methodology (refer §3.3.2).

The domain-specific and domain-application levels are the "usable" section of the KB and specific coral reef systems and hypotheses are defined here. Individual hypotheses are populated with domain-specific and hypothesis-specific data in the usable domain-specific and application ontologies. Because the usable ontologies employ the reusable domain ontology, the data is effectively separated from the KR structure and the generic restrictions on the properties of the reusable KB are unaffected.

To illustrate, two research hypothesis scenarios are presented. The first scenario involves a researcher who is searching for the cause of coral bleaching and is hypothesising about the correlation between light quanta levels, pH and high SST at a particular location on the GBR (e.g., Davies Reef). The "Coral Reef" ontology, which collectively encapsulates the hierarchy of generic domain ontologies, is imported to the "GBR" domain-specific application ontology (Figure 3.11) and populated with data relevant for that hypothesis (i.e., sensor data fed from the GBROOS Davies Reef weather station). The second scenario involves research conducted on the Moorea Islands in the French Polynesians by the Coral Reef Environmental Observatory Network (CREON)[30], the research hypothesis would employ the same generic "Coral Reef" ontology, but is populated with data pertinent to Moorea. Another domain-specific application ontology is created (Moorea_Island.owl) that imports the reusable ontologies and is populated with data, such as SST and Par, for the Moorea reef system (Figure 3.11). The elements and classes of the generic ontologies will be the same for either the GBR or Moorea Island location, but the instance data and rules that represent the hypothesis will differ.

[30] http://www.coralreefeon.org/

3.4.9. The Application Level – The Usable KB – The Inference Rules

The application ontology level is the final level of the KB hierarchy and where the hypothesis-specific data and inference rules are introduced. The application ontologies are singular in purpose, to implement the inference rules of a hypothesis. They import a populated domain-specific ontology (e.g., GBR.owl, Moorea_Island.owl, etc.), which contains the instance data and the lower "reusable" domain ontologies. Then, propositional testing is implemented through SWRL inference rules to perform tasks. Example tasks are to pose a hypothesis of the KB or to query the KB.

SWRL manages inference through Horn-like rules, which is a subset of predicate logic (FOL) and orthogonal to description logic (Antoniou et al. 2001). A SWRL inference rule is atom centric and contains antecedents (the body) and consequences (the head). The antecedent (body) of the rule represents the information supplied in order to draw a conclusion, and the consequence (head) is the implication that is ultimately drawn (Horrocks et al. 2004). A SWRL rule has the form:

$$\text{Body}(X_1 \wedge X_2 \wedge X_n) \rightarrow \text{Head}(Y_1 \wedge Y_2 \wedge Y_n)$$

Both the body and the head can consist of conjunctions of atoms (X and Y). These atoms can be in the form of an OWL class, an OWL property or a declaration of *owl:differentFrom* or *owl:sameAs* that refers to OWL individuals or OWL data values.

The SWRL rules are written to represent the hypotheses posed by a marine researcher. Many hypotheses can be fashioned in a Horn clause form due to the syllogistic format of propositional logic. The propositional logics are used to design deduction and induction rules and, in future development, reactive or production rules. An example of an inference rule to determine whether a bleach warning should be issued, based simply on the SST, would be:

```
    Coral_Reef(?z) ∧ Coral(?x) ∧ hermatypic(?x,true) ∧
Sea_Surface_Temperature (?z, ?y) ∧ swrlb:greaterThan(?y, 32)
                  → bleaching(?x, true)
```

When translated the antecedent (body) states the conjunction between all factors included in the bleaching process. Explicitly, if the "hermatypic" Boolean data-type property of an individual (x) in the "Coral" class is true and if the Coral Reef (z) to which the individual belongs has an SST (y) above 32, the result would be inferred (concluded) automatically. In this case, the

resulting consequence (head) will automatically change the "bleaching" data property of the "Coral" class individual to the Boolean value "true".

Autonomy in the propositional analyses processing was the motivation for enlisting both OWL-DL and SWRL inference. SWRL inference rules work in conjunction with DL. In contrast to DL which reasons over the ontology classes, SWRL reasons about individuals as members of classes and not the classes themselves. The classification of the KB's class and individual structure and subsumption is accomplished via the reasoning engine and the OWL DL constructs. Then, the SWRL rules are passed to the Jess inference engine (Jess 2006) for inference over the ontology instances, directly or using the "SWRL to Jess bridge" component of Protégé (O'Connor et al. 2005).

The "usable" application ontology level takes advantage of the OWL modularity support to allow for multiple hypotheses of one instance of the KB. Modifications can be easily made to the rules of an application ontology or the interchange of instance data for a different proposition. To alter the hypothesis in question, the only changes required entail either simple modifications to the body and head of the inference rule or, if needed, the creation of a new application (rules) ontology. The instance data and lower ontologies are not affected. In fact, as the enquiries themselves change, many application ontologies could be applied to the same instance of the KB for multiple hypotheses.

3.5. Justifications

Semantic technologies present data and information in new ways to the computer for processing. These technologies have great potential to fully automate data integration and knowledge generation processes. The Semantic Reef model is a proof-of-concept to apply these technologies to data such as decidability and inference and to manipulate the data in new ways to create knowledge. The technologies offer different techniques for bridging disparate data sources, which in this case is environmental data for marine research, both historic and current (and growing).

Clearly reefs are highly complex, interdependent ecosystems. To describe an ecosystem in a singular large ontology would not have been simple to create, implement nor maintain because the multi-scale relationships are complex. Therefore, an intra-ontological and inter-ontological design was adopted. The ontologies within the project were developed firstly using an intra-ontology design methodology where each node of the expert's model (Figure 3.1) was represented

in ontological form. Then, an inter-ontological methodology was implemented to take advantage of OWL's support for importation and reusability.

The inter-ontology design strategies encompassed decisions about which type of ontology was most appropriate to describe a concept. The decisions were based on the need for efficiency, reusability and flexibility. Efficiency and reusability was attained by applying complexity only when it was appropriate or warranted to describe a concept or achieve a purpose. An ontology should remain straightforward where possible, such as a lightweight taxonomy or informal ontology (i.e., written using OWL-Lite) to maintain efficiency and lower computational work by the reasoning engine. If a concept warrants complex relational definitions, then the flexible modular design allows for more formal ontologies to be created. The formal ontologies were a more appropriate choice to describe complex concepts because they offer constructs for describing intricate, multi-scale relations.

The ontologies were designed as separate standalone files that can be imported to the higher levels of the ontology base. The lightweight ontologies were written in OWL-Lite to reduce complexity and reasoning time and foster reusability. The relationships can be modified and classes can be added or removed for individual ontologies. The hierarchy starts with the taxonomies and informal ontologies at the base level, which are imported to the heavyweight formal domain ontologies. Although finer granularity in property and relationship restrictions and definitions is applied at this level, the concepts are still general enough to be reusable for any coral reef.

The domain-specific and application ontologies at the highest layers import the reusable ontologies to effectively maintain the usable components of the KB. The domain-specific ontologies define explicit details of a particular reef, including instance data relevant to that reef. Then, at the highest level of the KB, application ontologies are created that import all lower ontologies and define the inference rules dependent on the hypothesis proposed. The effect is a separation of instance data from the inference rules which creates flexibility when posing questions of the data. Therefore, simple modifications to the hypothesis, or entirely new application ontologies, can be created dependent only on the researcher's specifications.

3.6. Summary

This chapter described the reusable and usable ontology design developed to express the concepts of a coral reef ecosystem in a form understandable, and as a result decidable, by a computer. The ontologies were engineered in a "ground-up" physical hierarchy. The informal

taxonomies and simpler concepts were imported into the higher level ontologies of the hierarchy and at each level higher, the relationships, and their axioms and restrictions, grow more complex. The ontologies serve a designated purpose at each level; the more complex the ontology the narrower the purpose. The modular design of the coral ecosystem KB is justified because the flexibility is diminished with the added complexity. Thus separate informal to formal ontologies were created to describe the simple to more complicated concepts. The strategy maintains scalability and reusability for future hypotheses independent of reef type, location and community makeup.

The domain-task and application ontologies at the "usable" level of the KB employ the general and domain "reusable" ontologies beneath. The questions are in the form of finely detailed inference rules written to the system, as propositions, to infer conclusions about a specific issue on any arbitrary reef, regardless of location. The ontology design strategy aimed for extensiveness, flexibility and reusability and thus, hypotheses can be posed of any coral reef simply by populating the KB with instance data pertinent to that locale. Once data is coupled to the ecological and environmental ontology base, queries and propositional tests through inference can be applied.

A reverse-hypothesis methodology was employed to validate the Semantic Reef KB. The process and positive outcome is described in detail in the following chapter.

Chapter Four

The Validation of the Knowledge Base

4.1. Chapter Synopsis

The development of the Semantic Reef KB was described in the previous chapter. The substantiation of the KB is detailed in this chapter.

To validate the KB, a reverse-hypothesis approach was taken. This approach involved comparing the inferred outcome from the KB to the historic events and the ensuing observational research; that is, to ground-truth the system. The Semantic Reef KB is an observational hypothesis tool, where the inferred outcome from the KB can be observed *in situ.* The accuracy of the KB as a tool for hypothesis-driven research is tested so the questions posed of the system in the future will infer *in silico* outcomes, which can be observed *in situ* to prove a hypothesis true or null.

The subjects used in the validation were the mass coral bleaching episodes that occurred on the GBR in 1998 and 2002 (Figure 4.1). The KB was populated with the historic SST data from previous research on the GBR. Then, the outcomes of the system's rules were evaluated against the

Figure 4.1 – Coral bleaching - Photo by Ray Berkelmans, AIMS.

historic data analyses and *in situ* field observations of these mass bleaching events.

The coral bleaching phenomenon and the current methods to monitor sea temperatures and assess the probability of bleaching events are discussed below. Then, the validation process is explained, including descriptions and justifications of the logics and rules formed to mimic the conventional coral bleaching prediction metrics. To conclude, the successful comparison between the historic analyses and the results from the KB inference rules is illustrated and discussed.

4.2. Background - The GBR and Coral Bleaching

The GBR is sometimes referred to as the single largest organism in the world. The GBR covers an area of 348,000 square kilometres and spans ~2300 kilometres of the Australian State of Queensland's coast from the northern tip to Bundaberg. It is the largest coral reef system in the world and is made up of many billions of tiny coral formations. Fishing and tourism activities in the GBR contribute significantly to Australia's economy (Access-Economics 2005). However, coral is being damaged by coral bleaching and the viability of the GBR is threatened. A causal factor of coral bleaching is high oceanic temperatures so global warming will, no doubt, increase the possibility of bleaching occurences (Hoegh-Guldberg 1999).

Corals live in a symbiotic relationship with single-celled dinoflagellates called zooxanthellae that live within the coral's tissue at extremely high densities. Zooxanthellae photosynthesise to provide an essential food source for corals and the photosynthetic pigments also provide corals with their brilliant colours. In exchange, the corals provide the zooxanthellae with a place to live.

Coral bleaching results from a breakdown of this symbiotic relationship caused by a stress conditions such as higher-than-normal sea temperatures (Brown 1997). Reef corals are very sensitive to sea temperatures outside their normal range. Elevated temperatures of 1^0 Celsius above the long term monthly summer averages are sufficient to cause the stress factors that result in coral bleaching in many coral species (Brown 1997).

Bleaching is the product of the breakdown in the symbiosis. The energy from the sun is normally used by the zooxanthellae to produce food. However, as the temperature increases the algae begin to produce oxygen radicals that are highly corrosive and damage both the zooxanthellae and thus the coral. At high temperatures the light is toxic to the algae so when temperatures exceed threshold levels for long enough, the symbiotic relationship between the zooxanthellae and the corals breaks down (Jones et al. 1998). The coral then expunges the zooxanthellae and then

bleaching results. Algae give corals their characteristic brownish colour and when they are expelled what remains is the white skeleton clearly seen through the corals transparent tissue. The white appearance is called bleaching (Jones et al. 1998).

If stressful conditions continue, the corals bleach and die. However, a bleached coral is not a dead coral and, if stressful conditions abate, corals can regain their zooxanthellae and return to their normal healthy colour (Marshall and Schuttenberg 2006).

Spatially extensive, or mass, coral bleaching events have been largely attributed to anomalously high temperatures. However, the extent of the bleaching is affected by a combination of many local factors: the community or nutrient composition, hydrodynamics (i.e., bathymetry,

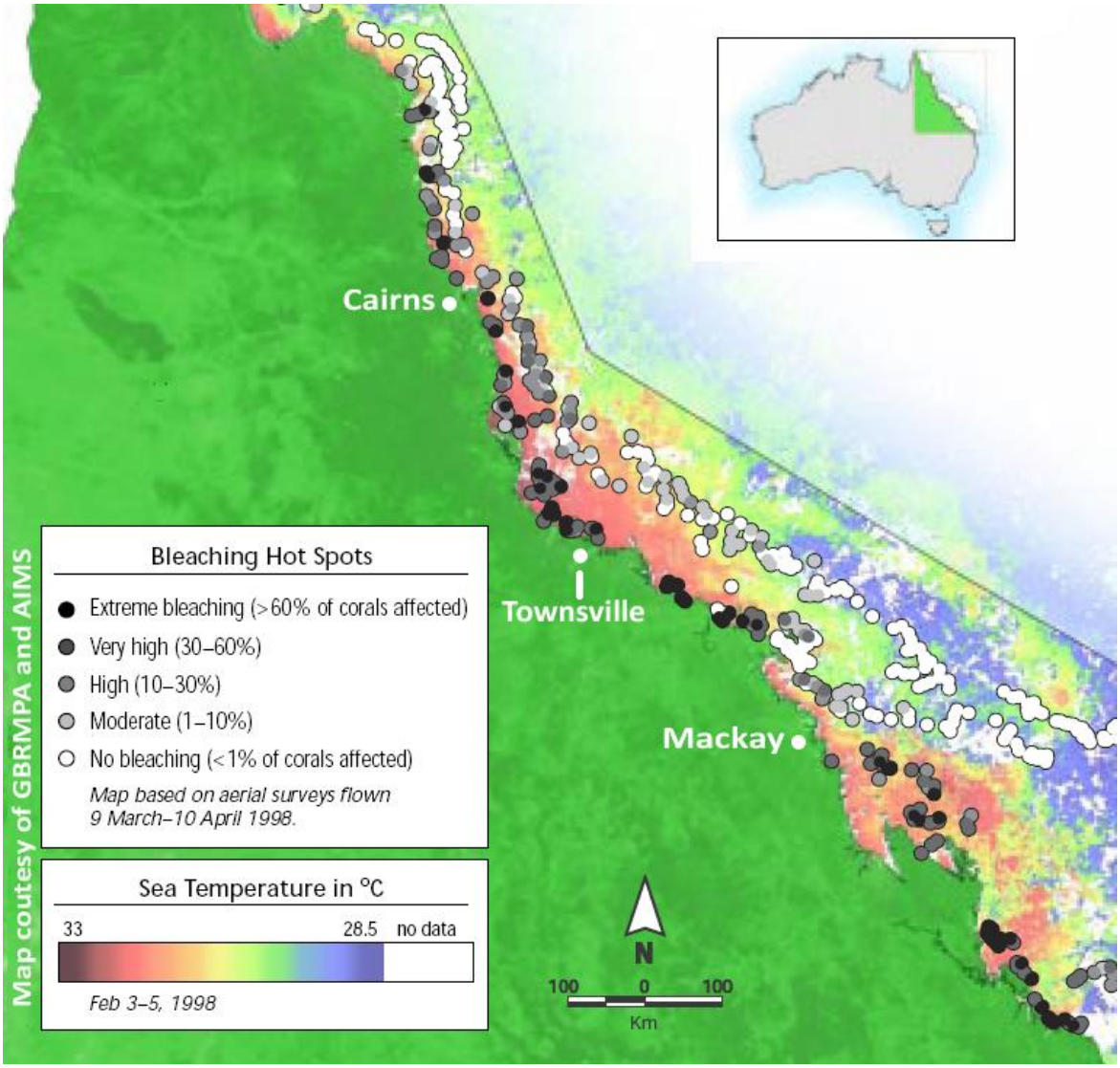


Figure 4.2 – Map showing bleaching on the Great Barrier Reef as seen from aerial surveys in 1998 (Berkelmans et al. 2002).

currents, tides, etc) and cloud cover. Other factors which have been suggested to influence bleaching include water depth, location, salinity, light intensity, pollutants, exposure, pH and even sedimentation (Hughes et al. 2003; Brown 1997; Marshall and Schuttenberg 2006). To accurately predict bleaching events, it is necessary to have a model which can assess all these factors.

4.2.1. Current Research Methodologies and Materials

Two major coral bleaching events occurred in the Coral Sea and on the GBR in the late

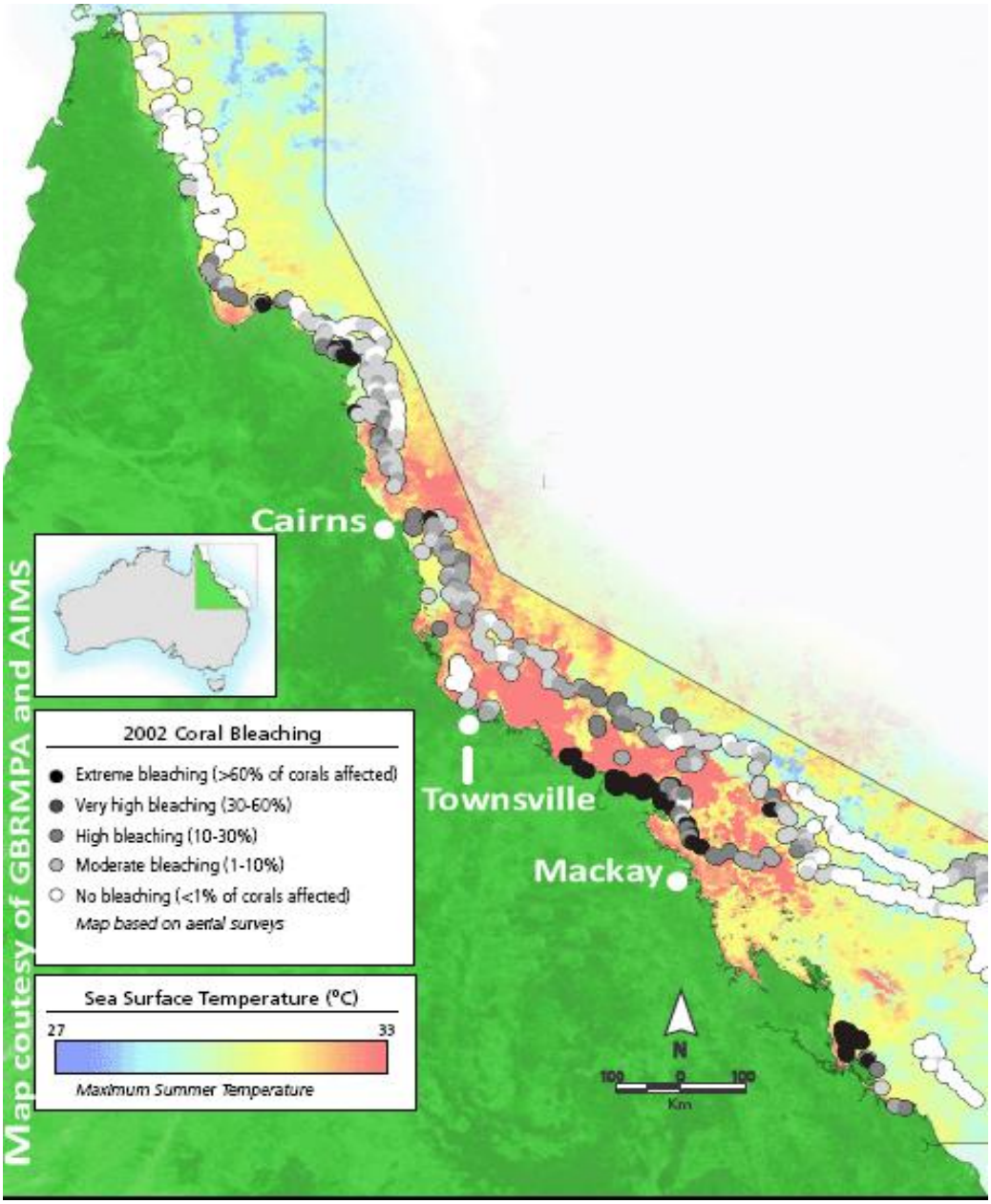


Figure 4.3 – Map showing bleaching on the Great Barrier Reef as seen from aerial surveys in 2002 (Done et al. 2005).

summer (February and March) of 1998 and 2002 (Figure 4.2 and Figure 4.3). Mild bleaching began in late January of the 1998 summer and intensified throughout February after hotter than normal temperatures. Every coral reef region in the world was affected by the 1998 bleaching event, which was the worst global bleaching event on record (Berkelmans et al. 2002). However, although the GBR suffered extensive bleaching during both events the summer of 2002 was more severe than the summer of 1998 (Done et al. 2005).

The bleaching severity during each event was assessed by underwater video survey at fourteen sites on the central GBR. These sites were selected to explore the relationship between accumulated thermal stress and bleaching severity (Berkelmans et al. 2004). Four of the initial fourteen sites were re-surveyed in 2002 to evaluate changes in the relationship between thermal stress and bleaching severity between the events. The four sites are located in the central section of the GBR: Kelso Reef, John Brewer Reef, Faraday Reef (via the Myrmidon Reef monitoring station) and Florence Bay at Magnetic Island (Figure 4.4). The reefs showed significant levels of bleaching in both the 1998 and 2002 bleaching events and have been used to estimate the relationship between accumulated thermal stress and bleaching severity (Gleeson and Strong 1995; Berkelmans 2002; Maynard, Anthony et al. 2008).

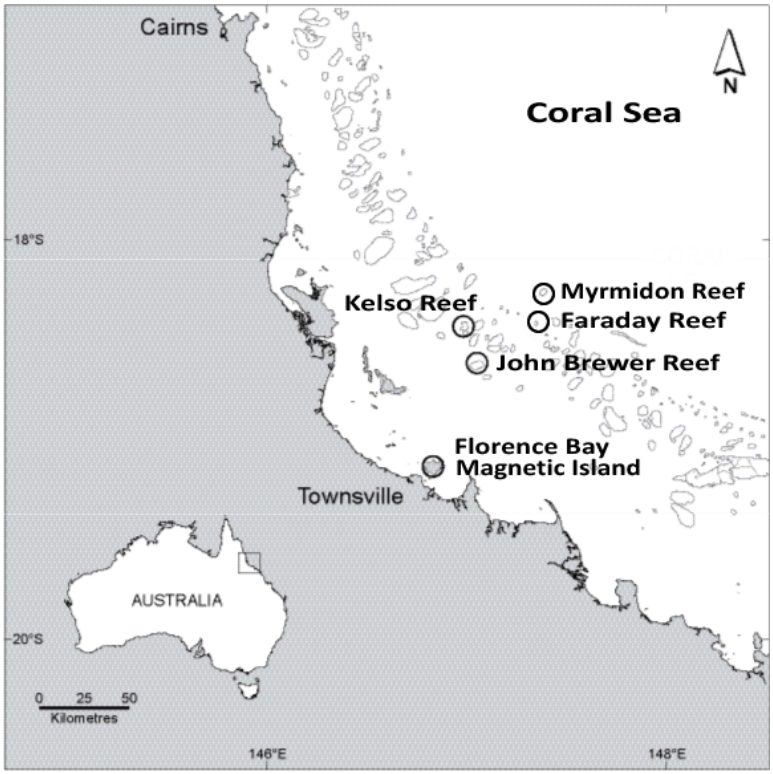


Figure 4.4 – Sitemap of the targeted reefs in this study.

The historical records of the temperature regime taken from the four reef locations were obtained by remotely sensed SST from the satellite platforms of the United States NOAA. The SST was measured by Advanced Very High Resolution Radiometer (AVHRR) and compiled for the central GBR region at ~4 kilometre resolution for 1992 to 2006 (Gleeson and Strong 1995; Maynard, Turner et al. 2008).

4.2.2. The SST Data

The data for the validation was supplied by the AIMS large-scale temperature monitoring program[31], which is in conjunction with the Great Barrier Reef Marine Park Authority (GBRMPA) and the Cooperative Research Centre (CRC) Reef (GBRMPA 2005; CRC 2006). Field surveys were conducted in 1998 by the Long-Term Monitoring Program (LTMP) team of AIMS and in 2002 by both the GBRMPA and AIMS. The raw data used in the surveys consisted of SST taken for the summer periods from Florence Bay and the Myrmidon, John Brewer and Kelso Reefs over a period of 8 years, 1995 through to 2003. The datasets contained minimum, maximum and mean daily SST for the reef slope area from 1995 – 2003 and the reef flat areas from 1998 - 2003 for all reefs except John Brewer Reef, where there was no available data for the 1998 summer.

Maynard's study (2004) of the thermal tolerance of major coral genera was the benchmark for the validation of the Semantic Reef KB. His research focused on four locations: Florence Bay and the Faraday, John Brewer and Kelso reefs. There was no *in situ* temperature data available for Faraday Reef, so the data for Maynard's study (2004) was extracted from Myrmidon and Dip reef monitoring stations and averaged to form representational temperature data for Faraday Reef. Representational environmental data is common practice in marine science due to the remote locations of the research subjects (Maynard 2004; Maynard, Anthony et al. 2008). The validation of the Semantic Reef system requires only corresponding outcomes from the data analysis; therefore, Myrmidon Reef was the focus for the tests here.

Of the datasets available it must be noted:

- There is no SST data for John Brewer Reef for the summer of 1998;
- The SST data was only collected from the flat area for the Myrmidon Reef prior to and including the summer of 1998;
- The SST data is not available for the summer of 1999/2000 for Myrmidon Reef; and

[31] The temperature data used here were made available via GBRMPA's "Long-term monitoring of sea temperatures" project (ID 133) and the Sea Monitoring Program database (coordinator, Ray Berkelmans). Loggers are placed on flat and slope regions of monitored reefs in the Northern, Central and Southern GBR.

- The flat temperature data was not available at Kelso Reef until the summer of 1997/1998.

4.2.3. Outcomes and Interpretations - Historical

The maps in Figure 4.2 and Figure 4.3 show the extent of bleaching severity for the 1998 and 2002 mass bleaching events, respectively. The four reefs are the focus of the validation and they are located in the Central GBR transect. The central region suffered high to extreme bleaching in the inner shelf reefs and moderate bleaching in the mid and outer shelf reefs during the 1998 event and moderate bleaching for inner, mid and outer shelf reefs during the 2002 (Maynard 2004; Maynard, Anthony et al. 2008). The SST graphs for the 1998 and 2002 summer periods for each site are shown in Appendix B and Appendix C (courtesy of Ray Berkelmans, AIMS LTMP).

4.2.4. Thermal Stress Indices for Coral Bleaching Analyses and Prediction

Coral bleaching risk is estimated by calculating thermal stress indices that measure bleaching severity based on temperature characteristics. Thermal stress indices, which detect thermal anomalies of the accumulation of heat stress, were applied in the historic studies of the 1998 and 2002 mass bleaching events. Four of these indices include the magnitude of SST anomaly (SST+), maximum summer temperature ($_{Max}$SST), the "HotSpot" anomaly and Degree Heating Days (DHD) (Berkelmans et al. 2004; Maynard, Turner et al. 2008; Maynard 2004).

The first metric, SST anomaly (SST+), calculates the temperature anomaly as the number of °C above the Long-term Mean Summer Temperature (LMST) observed at each site for that month and is ranged from +0.1°C to +5°C.

$$SST+ = SST - LMST$$

In contrast, the $_{Max}$SST is based on the Local Mean Summer Maximum (LMSM) temperature.

$$_{Max}SST = SST - LMSM$$

The "HotSpot" index was published in 1997 (Strong et al. 1997) and is also an anomaly metric. It differs from the previous metrics because it is not a typical climatological SST anomaly that is based on the average of all SSTs. Instead, the HotSpot anomaly is based on the climatological mean SST of the hottest month for the region, referred to as the Maximum Monthly Mean (MMM), (NOAA 2009b). Corals can stress when temperatures exceed only 1 °C above the

summertime maximum. The HotSpot metric is calculated as the difference between the average daily SST and the MMM SST:

HotSpot = SST – MMM

Only positive values are expected since the HotSpot is designed to show only extreme circumstances of thermal stress (NOAA 2009b).

The DHD index describes the accumulation of thermal stress. One DHD is calculated as one degree above the local LMST for one day. Two to Five DHDs are similar in calculation, where two DHDs are either two degrees above the local LMST for one day, or one degree above LMST for two days, and so forth. The DHD value is the summed positive deviations of daily average SST from historical LMST (CSIRO 2007) and is calculated as:

DHD = $\sum$ (IF (SST – LMST) > 0)

The Sea Surface Temperature plus (SST+), $_{Max}$SST, HotSpots and Degree Heating Days (DHDs) have been shown to be well correlated with the severity of bleaching responses during the 1998 and 2002 bleaching events (Berkelmans et al. 2004; Maynard, Turner et al. 2008; Liu et al. 2001).

The anomaly indices (SST+, $_{Max}$SST and HotSpot metrics) and the accumulation index (DHD) are the focus in the validation of the KB. The indices were back calculated for all the 1998 and 2002 survey sites for which temperature data was available and related to the severity of bleaching responses at those sites (Maynard 2004). Logical inference rules, DL and queries were used to mimic the aforementioned metrics and subsequently executed. The results from the logical inference related closely to those of the previous research results on the tolerance of corals to temperature changes.

4.3. The Validation Ontologies and Workflow

A task of the validation process was to create sets of computer-understandable axioms and inference rules to characterise the concept of coral bleaching. Specifically, known circumstances such as the rising SST and coral characteristics were explicitly defined in the rules and definitions of the domain-specific and application ontologies. A characteristic to describe coral is its type; it may be either hermatypic or ahermatypic. Hermatypic describes coral that contain and depend upon zooxanthellae for nutrients and is thus susceptible to bleaching (Schuhmacher and Zibrowius 1985).

In the Semantic Reef system the computer infers whether an event has occurred by using semantic definitions to describe the common bleaching indices with axioms and rules. The first tests use the anomaly metrics (SST+, $_{Max}$SST and HotSpots) defined as rules to infer possible bleaching. The final test uses the cumulative metric DHDs and is depicted by the SWRL query language (SQWRL pronounced "squirrel") for each reef (O'Connor et al. 2007). The rules and queries detect the temperature regime circumstances that indicate an event and are verified by the outcome of the Jess inference engine. The outcome was subsequently compared to the actual observed events from that time period.

4.3.1. The Domain-Specific GBR Ontology

Ontologies at the domain-specific level include the information about a particular reef, which in this case is the Great Barrier Reef. The hierarchy of reusable ontologies encapsulated within the "Coral Reef" ontology (as described in Chapter 3) is imported to a "GBR" domain-specific ontology (refer to Figure 3.2). To support modularity and to increase reusability and flexibility, ontology engineering commonly separates the data (instances) from the definitions (classes and roles) (Baader et al. 2007). This separation allows the ABox parts (instance data) to be changed without affecting the TBox parts (i.e., definitions generic to all reefs), which can be reused for DL reasoning (refer to Chapter 2 §2.4.1.5.6). Here, the data instances (ABox) are introduced at the domain-specific level to the "GBR" ontology and includes specific details of the individual reef, such as longitude, latitude, environment variables, etc. thus remaining separate from the "Coral Reef" domain ontology (TBox).

A GBR domain-specific ontology was created to define GBR related concepts. The separate reefs of the GBR were declared as new classes of the "GBR" ontology instead of single instances of a generic reef type. As separate reef classes they can be populated with temporal instances to infer knowledge based on the environment of that reef at a specific moment of time. The temporal reef instances are specialisations of that class by the different values of its data-type properties (i.e., the environmental and time variables). New classes were added to label reefs within the GBR system, among them were the reefs used for the validation (i.e., Myrmidon, John Brewer and Kelso reefs and Florence Bay). The individual reef classes were asserted as subclasses under the "Inner Shelf", Mid Shelf" or "Outer Shelf" classes depending on their geographical locations (refer Figure 4.5). In turn, the shelf location classes are sub-classes of the generic "Barrier Reef" class which, in turn, is one of a number of classes to define a generic type of reef (e.g., fringing, barrier, atoll, etc.).

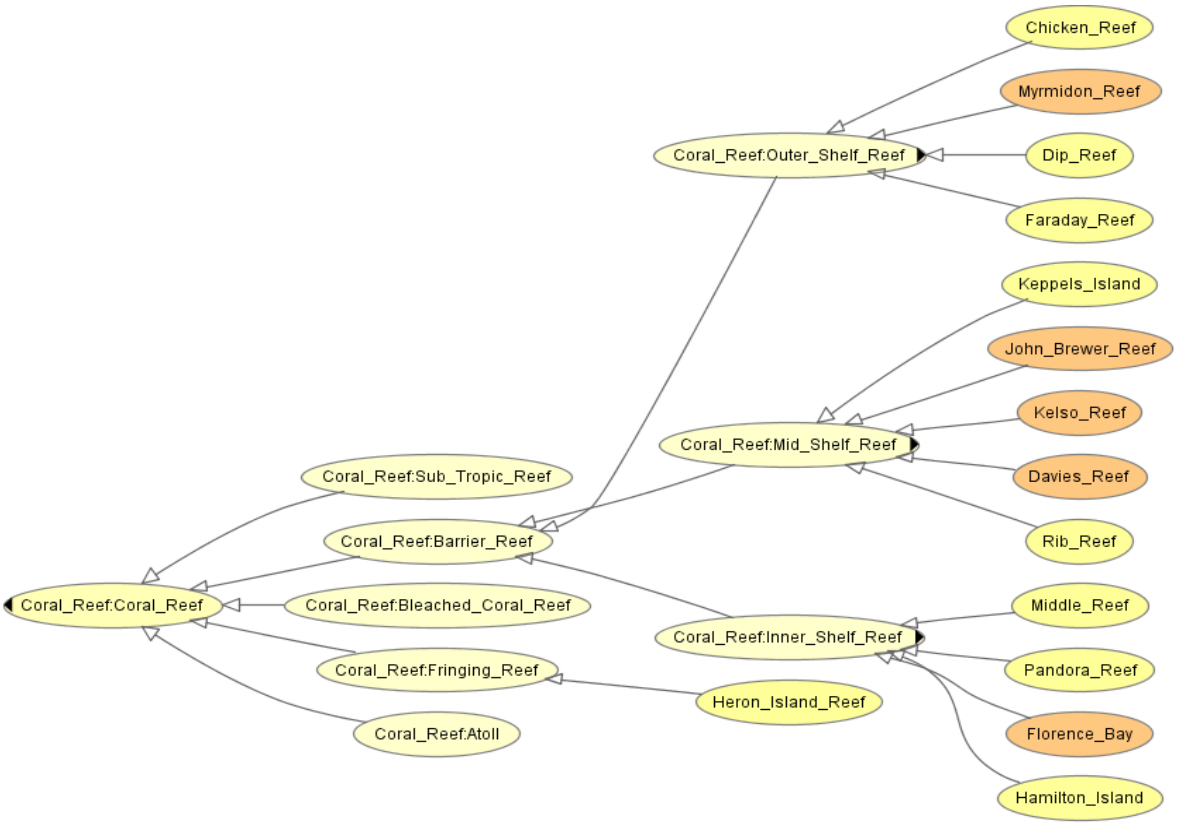


Figure 4.5 – A segment of the Coral Reef GBR ontology depicting the modular class structure.

The instance data is introduced at the domain-specific layer for inferring knowledge at the application (rules) layer of the KB. SST data from both reef flat and reef slope were used in the preliminary tests, which resulted in two instances per day for each reef. The technique to average the daily flat and slope temperatures of each reef as representational data is common practice for anomaly and cumulative metrics and was adopted here (Maynard 2004; Holmes 2008a).

4.3.2. The Application Ontology – The Inference Rules

SWRL rules are written for the different hypotheses at the highest layer of the ontology hierarchy, the application ontology. SWRL rules were created to mimic the anomaly metrics to infer a bleaching event and SQWRL was used to query the KB to mimic the accumulation metrics. Because the instance data is separated from the rules reuse of the KB was possible. Unlike DL, SWRL reasons about individuals as members of classes, not the classes themselves (Horrocks et al. 2004).

Propositional logic is orthogonal to DL and SWRL uses propositional logic, written as Horn rules, to reason over OWL individuals and infer new knowledge about these individuals (Baader et al. 2003). Specifically, reasoning engines apply DL when reasoning over a KB to classify and untangle classes and individuals. In contrast, SWRL inference rules are posed of individuals when the desired outcome is an inferred syllogistic conclusion. Therefore, when additional expressivity was required, either to pose a question of the KB or to infer knowledge into the KB, SWRL rules were added. The SWRL rules are passed to the Jess inference engine (Jess 2006) via the "SWRLJessTab", which is a SWRL to Jess bridging component of Protégé (O'Connor et al. 2005).

Non-monotonic rules are not permitted in SWRL at it only supports monotonic inference (Horrocks et al. 2004). Non-monotonic rules, also classed as defeasible logic, are true until new knowledge can prove otherwise or are "defeated" by other rules. (Antoniou et al. 2001). In contrast, a monotonic rule draws a conclusion that remains valid even after new knowledge is formed. For example, a rule may specify "a coral that is ahermatypic will not bleach". The system can then deduce any susceptible areas for bleaching based on that rule and the community coverage of a reef.

The OWA is both a limitation and an advantage of the Semantic logic systems. SWRL rules cannot be used to modify information in an ontology, for instance, removing a value or updating a current value (commonly required for aggregation). Hence, if a SWRL rule modifies an axiom that is currently defined in an ontology, non-monotonicity would ensue (O'Connor et al. 2007). Operational types such as aggregation are not DL-safe constructs, due to the OWA. They require binding individuals to unknown values which results in undecidability and therefore cannot be used as antecedents in SWRL rules. For example, a rule to average all values of SST contained within the KB would not return a true number, according to the open world regime of semantics and DL, because the system would always assume there could be more members with a SST value. Hence, the aggregation of a finite group of property values is not possible in an monotonic system, because the OWA can have an infinite number of members (Motik et al. 2005).

SQWRL is a query language that operates over OWL forms to query OWL ontologies and is based on the notion of DL-safe rules (Motik et al. 2005; O'Connor et al. 2007). On many occasions a monotonic or deductive logic system cannot handle all reasoning assessments because knowledge of a concept may not be complete. SQWRL queries are used when aggregation or selecting and counting functionality is needed. Further, SQWRL queries are not independent of

SWRL rules they function in conjunction with the rules and retrieve knowledge that has been inferred by the rules (O'Connor et al. 2008).

Individual SWRL rules and SQWRL queries were created in the "GBR Rules" application ontology to simulate the bleaching indices and infer bleaching events. The "GBR Rules" ontology imports the "GBR" domain-task ontology, which imports all lower ontologies and is already populated, via the workflow, with the relevant instances of the reefs in question.

The validation was successful because the outcome from the SWRL rules and SQWRL queries matched the physical outcomes from the historic bleaching events. The rules and queries, shown as examples in the following sections, followed the human readable syntax described in the W3C member submission for the SWRL standard (Horrocks et al. 2004) (which is supported in the Protégé "SWRLJessTab").

4.3.3. The Scientific Workflow

The temperature logger data was stored in tabular Comma-Separated Value (CSV) format and ported to the KB via a Kepler workflow. A scientific workflow was created to physically manipulate the data in preparation for the KB, to test the system's ability to provide a coral

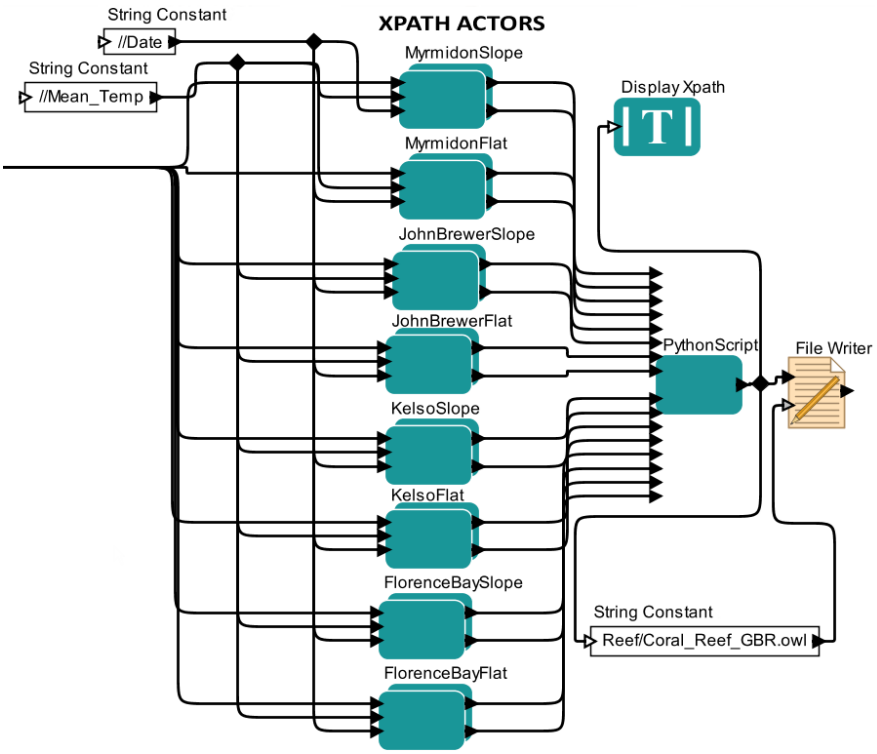


Figure 4.6 – XPATH actors in Kepler extracting temperature and date from each site.

bleaching "alert" (Figure 4.6). The workflow imports the data in XML format and, through the XPATH[32] actors in Kepler, extracts each date with its corresponding mean, minimum and maximum temperature data value. The workflow also extracts the LMST, LMSM and the MMM, which are components of the bleaching metrics.

The domain-specific ontology at the application level of the KB consists of classes that are designated named reefs within the GBR system, including the four focus reefs. All values from the workflow are passed to the KB to populate the environmental and temporal properties of each of the reefs. These properties are data-type values such as SST and date/time that are asserted to the instances created under each reef class, one instance per temporal value.

After the population process was completed, the KB contained 2579 reef instances. Of these reef instances, 696 were asserted to the "Florence Bay" class, 631 to the "Myrmidon Reef" class, 541 to the "John Brewer Reef" class and 711 to the "Kelso Reef" class. The environmental (SST, LMST, etc.) and temporal (date, time) values that were required for the validation exercise were asserted to each reef instance, as was the basic coral community makeup of the reef which includes coral species that are both zooxanthellate and azooxanthellate.

4.4. The Validation Tests and Results

4.4.1. The SST+ Index

The SST+ anomaly was tested by building an intricacy of SWRL rules to analyse the data. The process involved creating rules to establish the SST anomaly for each Coral Reef instance and categorising those that may be at risk of bleaching. The five SST anomaly categories range in degrees of risk directly related to the SST anomaly value, scaling from the lower risk of category 1 (+0.1°C to +1.0°C) to the higher risk at category 5 (+4.1°C to +5.0°C) (Maynard, Turner et al. 2008). Five sub-classes were created in the "GBR" domain-specific ontology under the parent "Bleached Coral Reef" class, to contain the temporal reef instances that were at risk: "Category 1 SSTplus 1" to "Category 5 SSTplus 5".

4.4.1.1. The SWRL Rules

The SWRL rules that mimic the SST+ metric infer reef instances to a bleach-risk sub-class that coincide with the anomaly categories. The rules automatically extract the following information of the reef instances from the KB: their date and daily SST values, the LMST for that

[32] http://www.w3.org/TR/xpath

location and the community composition. The workflow derives the LMST from the SST datasets for the 1997 to 2003 summer periods and from the community composition of corals from Maynard's (2008) research for the four locations. The workflow then dynamically asserts the values to the respective reef instance. The coral species are instances within the base level "Reef" ontology that are linked to the various coral reefs within the KB using the "hasPart" OWL object property of the "GBR" domain-specific ontology.

The antecedents for each rule test whether factors that relate to a bleaching occurrence are true. The factors include the presence of a positive SST anomaly and corals that are prone to bleaching (hermatypic). Firstly, the anomaly is calculated through the rules by subtracting the LMST from the daily average SST and it is then designated to an anomaly range, which coincides with a bleach watch category. For example, to infer a category 2 anomaly, the daily SST would be between 1 - 2°C higher than the LMST for that area, hence the rule will check for that particular range. The rule written in SWRL to determine the category 2 SST+ class is as follows:

```
                Coral_Reef:Coral_Reef(?x) ∧
       Coral_Reef:hasDailyAverageSSTof(?x, ?meanTemp) ∧
Coral_Reef:hasAverageLongTermSeaSurfaceTemperatureOf(?x,?LMST)
           ∧swrlb:add(?LMSTaboveStart, ?LMST, 1) ∧
           swrlb:add(?LMSTaboveCeiling, ?LMST, 2) ∧
    swrlb:greaterThanOrEqual(?meanTemp, ?LMSTaboveStart) ∧
        swrlb:lessThan(?meanTemp, ?LMSTaboveCeiling) ∧
             Coral_Reef:hasPart(?x, ?partCoral) ∧
                Reef_Stock:Coral(?partCoral) ∧
           Trophic:is_Hermatypic(?partCoral, true)
               →  Category_2_sstplus_2(?x)  ∧
                Coral_susceptible(?partCoral)
```

When the Jess inference engine runs, if all antecedents of the rule are satisfied, it will deduce which instances belong to the specific bleach-risk sub-classes.

4.4.1.2. The SST+ Index Results

The Jess inference engine returned the expected results and the instances with a positive anomaly were inferred to the correct bleach-risk category (Figure 4.7). The results from the total 2579 instances were as follows:

- 1016 instances had been inferred to the different bleach-risk classes for all summers from 1997 to 2003 (Figure 4.7);
- Of the 1016 instances, 754 were inferred to SST+ category 1, 229 to category 2 and 19 to category 3;

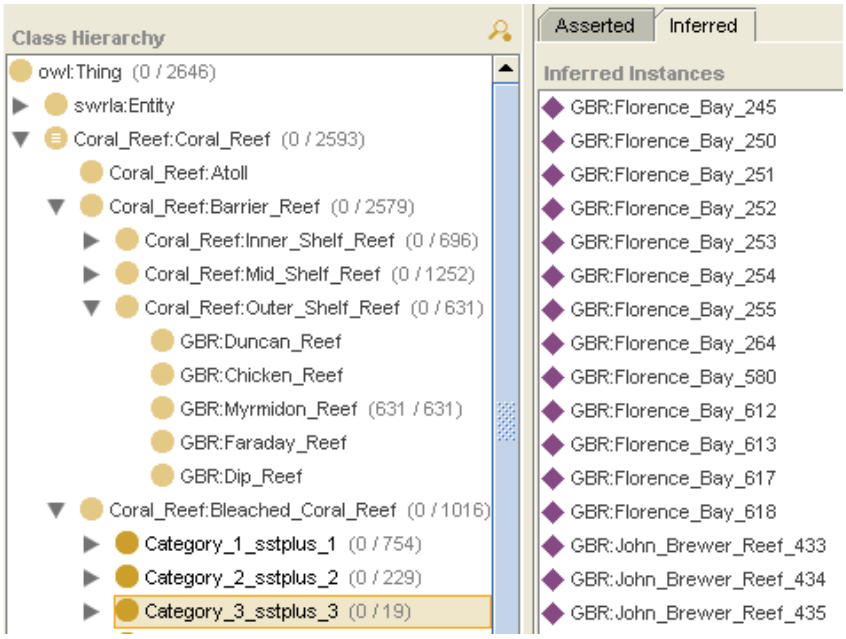


Figure 4.7 – The SST+ rules result in correct assertions and inferences – categorising the bleach alerts by SST+ categories.

- In the 1998 summer there were 77 instances inferred to SST+ category 1, 70 to category 2 and 8 to category 3; and
- In the 2002 summer there were 193 instances inferred to SST+ category 1, 114 to category 2 and 11 to category 3.

4.4.2. The $_{Max}$SST and HotSpot Indices

The $_{Max}$SST metric and the "HotSpot" metrics are anomaly based indices, similar to the SST+, therefore the rules to mimic the metrics could be devised in a similar fashion. That is, sub-classes were created, under the parent "Bleached Coral Reef" class to act as categorical bins for instances that prove true to the rules. The rules for the $_{Max}$SST and HotSpot indices calculate the anomalies and then infer all instances that fall into a specific category be placed in the correct sub-class.

4.4.2.1. The $_{Max}$SST and HotSpot SWRL Rules

The SWRL rules establish the $_{Max}$SST and HotSpot anomalies for each coral reef instance. The resultant anomalies are different values for $_{Max}$SST and HotSpots because the climatologies differ. The climatology for the $_{Max}$SST is LMSM whereas HotSpots are derived from the MMM

climatology (NOAA 2009b). Because the basic structure of the formulae is similar the rules are interchangeable with the exception of the climatology value.

Initially, the SWRL rules were created separately for both indices to confirm accurate outcomes. The inferred outcome for the HotSpot metric should return the smallest amount of temporal reef instances. Because the MMM is the highest climatology and would normally produce mostly negative values, the HotSpot metric is designed to show only instances where the SST is extremely high and conducive to bleaching. In contrast to the Hotspot metric, the $_{Max}$SST metric results should range between the HotSpot and the SST+ (category 2) metric results due to the LMSM climatology it employs. The LMSM is derived from the average maximum summer temperature over a period of years, which is higher than the LMST but mostly lower than the MMM. Therefore, the outcome from the inference rules should find the instances extracted as $_{Max}$SST will factor between the HotSpot instances and the instances inferred to the SST+ "category 2" sub-class.

4.4.2.2. The $_{Max}$SST and HotSpot Indices Results

Graphical representations of the results for the anomaly indices are shown in Appendix D (1998) and Appendix E (2002). The results for rules which mimic the $_{Max}$SST were as follows:

- A total of 687 reef instances were inferred to the $_{Max}$SST "category 1" class (for all summers from 1997 to 2003);
- Of the 687 inferred reef instances, 131 were for the 1998 summer and 270 for the 2002 summer;
- One instance was inferred to the $_{Max}$SST "category 3" class (Florence Bay for the 2002 summer period).

The HotSpot rules also showed similar correlation to previous research as follows:

- A total of 151 reef instances were inferred to the HotSpot "category 1" class (for all summers from 1997 to 2003);
- Of the 151 inferred reef instances, 12 were for the 1998 summer and 29 were for 2002 summer;
- One instance was inferred to the HotSpot "category 2" class (Florence Bay for the 2002 summer period).

To indicate a level of severity, rules were created to test all three anomaly indices simultaneously. The instances disclosed with these rules were the worst cases of thermal stress. The SWRL rules to extract any instance with all three anomaly indices satisfied for the 1998 summer period is:

```
Coral_Reef:Coral_Reef(?x)  ∧
Coral_Reef:hasDailyAverageSSTof(?x, ?meanTemp)  ∧
Coral_Reef:hasLongtermMeanMAXSummerSSTOf(?x, ?LongMax)  ∧
swrlb:greaterThan(?meanTemp, ?LongMax)  ∧
Coral_Reef:hasMAXMonthlyMeanSSTOf(?x, ?MMM)  ∧
swrlb:greaterThan(?meanTemp, ?MMM)  ∧
Coral_Reef:hasAverageLongTermSeaSurfaceTemperatureOf(?x,?LMST)
∧  swrlb:greaterThan(?meanTemp, ?LMST)  ∧
Coral_Reef:hasDateOf(?x, ?date)  ∧
temporal:during(?date, "1997-12-01", "1998-03-01")
→ All_1998(?x)
```

The results from these rules had a distinct correlation to the historic bleaching events (shown in Appendix D and Appendix E):

- The inferred results for the 1998 summer were as follows:
  - There were 3 days at Florence Bay (8^{th} - 11^{th} February) disclosed, and
  - 9 days at Kelso Reef (13^{th},14^{th} and 16^{th} – 22^{nd} February)
- The inferred results for the 2002 summer were as follows:
  - There were 11 days at Florence Bay (6^{th} – 8^{th} January and the 2^{nd}, 7^{th} – 9^{th} and 11^{th} – 14^{th} February),
  - 7 days at Kelso Reef (3^{rd}, 4^{th} and 9^{th} – 13^{th} February),
  - 4 days at John Brewer Reef (10^{th} – 13^{th} February), and
  - 5 days at Myrmidon Reef (5^{th} – 8^{th} February).

4.4.3. The Degree Heating Days Index

Because the DHD metric entails aggregation to sum the summer anomaly values, SQWRL queries were required. The queries were syntactically the same for each reef with the exception of the summer period requested. An example of the SQWRL inference query is written as:

```
GBR:Myrmidon_Reef(?x)  ∧
Coral_Reef:hasDailyAverageSSTof(?x, ?sst)  ∧
```

```
Coral_Reef:hasAverageLongTermSeaSurfaceTemperatureOf(?x,?LMST)
        ∧  swrlb:subtract(?anomalyFB, ?sst, ?LMST)  ∧
           swrlb:greaterThanOrEqual(?anomaly, 0)  ∧
              Coral_Reef:hasDateOf(?x, ?date)  ∧
   temporal:after(?date, "1997-11-30", temporal:Months)  ∧
    temporal:before(?date, "1998-03-01", temporal:Months)
                  → sqwrl:count(?x)  ∧
                   sqwrl:sum(?anomaly)
```

This particular query calculates the DHD for Myrmidon Reef for the 1997/1998 summer period.

4.4.3.1. The DHD Index Results

There was a distinct match to the historic research in the outcome from the SQWRL rules. Appendix F and Appendix G show the DHD graphs for 1998 and 2002 for each reef, respectively, and depict both the lineal DHD from the actual research (blue line) and the result from the inference query (red line). In comparison to other summer periods, the results indicate high DHDs for the 1998 and 2002 summer periods (Table 5.1). In the 1998 summer period, extremely high DHDs were apparent for Florence Bay, which coincides with the historical research and with the bleaching occurrence at that reef. In the 2002 period the SQWRL rules returned extremely high DHDs for all four sites, which again matched the historical observations.

	Summer Degree Heating Days (DHD)						
Location	***1997***	***1998***	***1999***	***2000***	***2001***	***2002***	***2003***
John Brewer Reef	10.1	NA	47.41	4.68	4.996	75.68	19.28
Kelso Reef	1.88	46.87	45.05	1.44	5.18	76.01	19.09
Myrmidon Reef	0.19	35.06	15.58	0	10.4	71.2	10.94
Florence Bay	26.59	82.38	47.63	10.55	3.17	77.47	7.13

Table 4.1 – Results from the DHD queries for all summer periods for each reef studied.

4.4.4. Overview and Discussion of the Inference Rules Results

To ground-truth the system, the coral reef instances inferred to belong to a particular bleach risk class were compared to the historic records on the bleach events. The successful correlation between the inferred results and the actual results proved the system to be accurate.

Except for the 2002 summer period, the results for the 1998 period showed above average high temperatures compared to the other years (Appendix D and Appendix F). The 1998 inferred results concluded Florence Bay and Kelso Reef both had SST+ category 1 to category 3

temperatures from late January to the end of February. The flat of Myrmidon Reef, which is an outer shelf reef, received category 2 SST+ from mid February to the end of the month (N.B., the temperature was not recorded for the Myrmidon Reef slope in 1998). Because there were no recorded temperatures for John Brewer Reef in the 1998 summer it was not possible to test and hence, there are no inferred bleach risk instances for this time. The outcome from the rules coincides with the bleaching occurrences observed for each reef, particularly Florence Bay and Kelso reefs (Figure 4.8) which showed signs of bleaching throughout late February in 1998 (Maynard 2004).

The 2002 inferred results showed rising temperatures were in the early to mid February period (Appendix E and Appendix G). The Florence Bay site had a significant rise in temperature throughout the 2002 summer period (SST+ category 1). From the 6th to the 10th of January the anomaly rose to category 2 and 3 SST+. However, the longest and hottest period was from January 30th to February 16th when the highest temperatures were reached. The Myrmidon, Kelso and John Brewer reefs were similar to Florence Bay, with rising temperatures (SST+ category 1) throughout

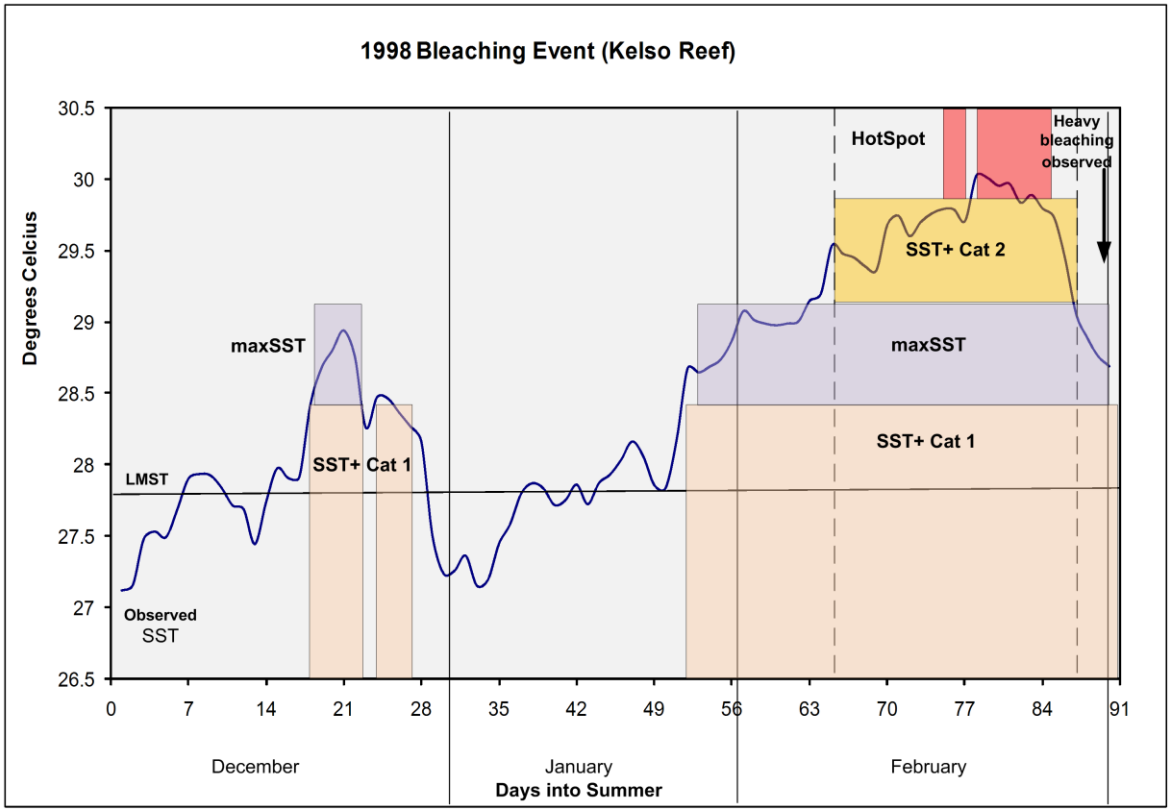


Figure 4.8 – SST data from Kelso Reef for the 1998 summer period (blue line) (GBRMPA 2005); rectangle overlays are regions that inferred a high risk of coral bleaching.

December and January, then moving to SST+ category 2 and 3 for the first half of February. These outcomes coincided with the observed bleaching occurrences for the 2002 period, when signs of bleaching began to show from January and the subsequent mortality rose towards the end of the summer period (CSIRO 2007; Maynard 2004).

The inference rules allotted the reef instances to be members of the correct bleach risk classes. For example, the Florence Bay instances from late January 1998 were sent to the lower SST+ category 1 bleach risk sub-classes. Then as the temperature intensified in mid February, and the highest temperatures for the summer occurred, the instances were inferred to the higher risk categories. The SST graph for Kelso Reef during the 1998 summer, with the inferred categories superimposed, is shown in Figure 4.8 and depicts the correlation between the inference rules outcomes for the SST+, $_{Max}$SST and the HotSpot metrics. Similar graphs for each reef are presented in Appendix D for the 1998 summer, and Appendix E for the 2002 summer and the correlating results are shown in each graph.

Because the SWRL inference rules and the SQWRL queries can accurately imitate the standard implementation of the bleaching metrics the system may be used to alert bleach risk areas. However, there are a number of initiatives currently in place to monitor and predict bleaching-related anomalies (Maynard, Turner et al. 2008; Hendee et al. 2008). The types of products developed by the NOAA, GBRMPA, CSIRO, and the Australian Bureau of Meteorology under the ReefTemp (CSIRO 2007) and ICON/CREWS (NOAA-ICON/CREWS 2008) initiatives make spatial predictions about bleaching severity. Many proposed monitoring and management measures are only effective given early warning systems, such as ReefTemp and ICON, that assess where bleaching and bleaching-induced mortality are likely to be most severe.

Currently the warning systems use the temperature regime and do not take into account the great number of variables that could also contribute to the bleaching phenomenon (Marshall and Schuttenberg 2006). The coral bleaching phenomenon is still being researched to determine why some coral species bleach in some areas but not in others and why some corals of the same genera, given time, adapt to higher SST (Hughes et al. 2003). Hence, the semantic ability will enable scientists to explore phenomena in new ways, because it can easily modify or add new relevant information to the ontologies such as chemical factors (e.g., nitrogen levels) or biological factors (e.g., fresh water flume occurrences). The Semantic Reef system is a hypothesis tool that can disclose these anomalies in the data.

4.5. Summary

The reverse-hypothesis (converse) approach was adopted to validate the Semantic Reef system. The ontologies in the KB described the ecology of a coral reef and the tolerance and interdependence of reef organisms like corals, to physical parameters (temperature) and the presence of zooxanthellae (hermatypic). Inference rules were created in the tests to mimic the stress indices relating to the SST anomaly and the accumulation of high SST. The indices included the SST+, $_{Max}$SST, HotSpot and the DHD metrics. The SWRL inference rules calculated the SST anomaly and automatically inferred the temporal reef instances to a SST+ category (ranging from 1 to 5), a $_{Max}$SST category (from 1 to 3) and a HotSpot category (from 1 to 3) for the summer periods from 1997 to 2003. Due to the OWA, SWRL does not support non-monotonic rules such as those that require aggregation. Therefore, SQWRL queries were employed to derive the DHDs for each reef because aggregation was necessary.

The outcome of the inference rules and queries was compared to the known monotonic facts from previous research of the 1998 and 2002 bleaching events on the Great Barrier Reef (Maynard, Anthony et al. 2008; Berkelmans et al. 2004; Hughes et al. 2003). The accuracy and consistency of the rules confirmed the validity of the system.

Although the stress factor most commonly associated with bleaching is elevated sea temperature, there are a number of other causal factors (Hoegh-Guldberg 1999). Additional stresses such as high light intensity, low salinity and pollutants are known to exacerbate coral bleaching (Berkelmans 2009; Maynard, Anthony et al. 2008). The Semantic Reef system expands upon the current bleaching research initiatives by providing a method to automate the data analysis and processing of disparate data-streams. The ground validation method proves the quantifiable accuracy of the prototype within the scope of a single hypothesis. The following chapter will demonstrate the differences semantic technologies combined with scientific workflows offer research via observational hypotheses.

Chapter Five

New Hypothesis Generation

5.1. Chapter Synopsis

The previous chapter described how the Semantic Reef KB was verified through a reverse-hypothesis methodology that compared the outcome of semantic inference against historic coral bleaching events.

This chapter expands on the substantiation exercise by illustrating the capabilities of the Semantic Reef system. In particular, the adaptability of the system for the discovery of new phenomena and the application of different hypotheses will be discussed. This discussion includes scenarios where the actual hypothesis is not apparent prior to gathering or sourcing the data and where flexibility is required of the system.

The data and rules from the previous chapter, which automated coral bleaching alerts based on historic SST data, are extended here in a series of demonstrations. The demonstrations show the benefits semantic technologies offer to hypothesis-driven research. Initially, the historic SST data is replaced by live data. The SST values are automatically mapped from near real-time remote sensor sources to temporal reef instances of the KB, via a Kepler workflow, to infer a coral bleach risk.

Data integration is demonstrated by introducing other factors such as Par and salinity to the KB. The other factors are incorporated in example observational hypotheses as antecedents of the inference rules to show the support the Semantic Reef system has for flexible hypothesis design and data integration. The focus of these example hypotheses is the unknown causal factors of the coral bleaching phenomena. Research into the causal factors suggests they may induce a coral bleaching event because bleaching is believed to result from an accumulation of contributory factors, not simply SST alone (Hughes et al. 2003). The current data on other potential causal factors is derived from separate autonomous sources. However, if the disparate data can be coupled with the greater range of environmental information new questions can be posed of the system. The propositions here are exemplar statements of the specific conditions of a hypothesis, where phenomena in the data are disclosed for *in situ* observation.

Finally, the benefits and value of fundamental functions of semantic technologies to hypothesis-driven research is demonstrated. The functions include automatic classification of concepts that can link latent connections automatically. Data representation is a technique in marine research used to apply data from one reef as indicative of other unmonitored reef systems. The types of reefs are concepts adopted in the data representation regime such as reef-types by location, by community make-up, by its biomass, etc. The reasoning engine can classify the KB via explicit assertions and axioms in DL to automatically link reef-type concepts.

5.2. The Semantic Application - Benefits and Distinctions

Flexible hypothesis design, data integration capabilities, automation and reuse are benefits and distinctions in the application of semantic technologies to hypothesis-driven research. Due to the unstructured nature of semantic KR, flexibility in hypothesis design is a possible functionality. Moreover, because a hypothesis can be structured and changed, independent of the data collection processes, bottlenecks in data analysis stages can be addressed. Data integration is a key function of semantic technologies because the OWA allows for new information to be easily added to a KB. Also, the latent links in data can be connected by inference, which is a powerful concept for the automatic extrapolation of new knowledge. These benefits are incorporated in the Semantic Reef KB which distinguishes it from other KR systems or hypothesis research methods.

5.2.1. Versatility in Hypothesising

Existing methods of scientific research may not scale as the deluge of data continues to grow. Current methods require the collection of data and the formulation and testing of hypotheses to be highly structured and contained as one action. However, with the rapid growth in data it would be beneficial to knowledge discovery to apply a more unstructured, organic methodology to hypothesis design.

A researcher using the Semantic Reef system is not required to predetermine the precise hypothesis prior to data collection and the population of the KB. Rather, the questions can be as flexible as the researcher requires; they may evolve as new data become available or as ideas grow and/or epiphanies emerge. For example, a researcher may initially propose a bleaching event with two factors (e.g., SST and salinity) and then decide to also include some unorthodox factors to the hypothesis. Other seemingly unconnected factors such as the sales of a particular fertiliser, the documented catches for a species of fish or a scheduled river dredging could be added to investigate their impact on inshore reef systems (Figure 5.1).

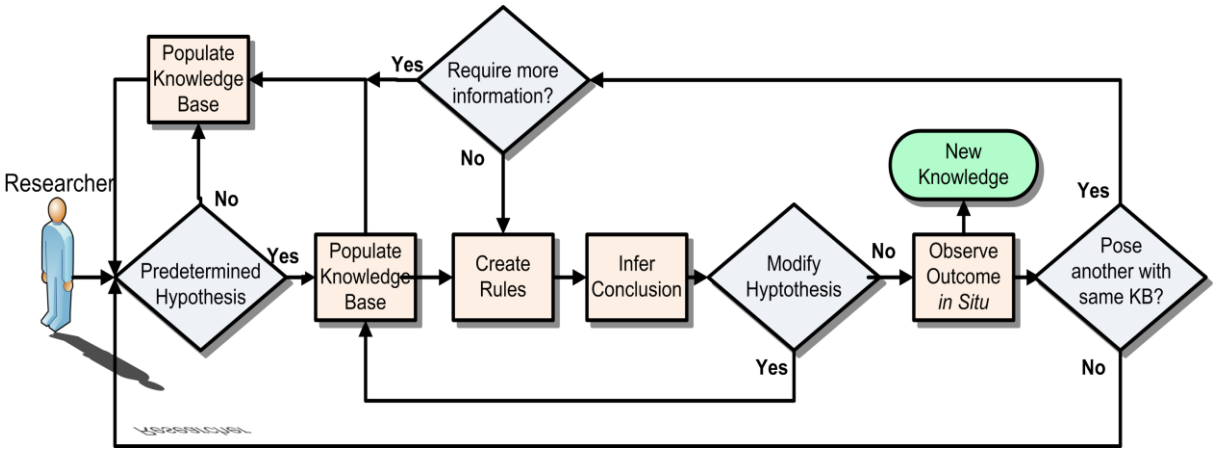


Figure 5.1 – A flowchart of the hypothesis design process. The propositions are fully flexible in light of new ideas or additional interesting data.

The types of questions that may be posed can be very versatile. If information is available and can be imported to the KB it can be added as a factor in any hypothesis. To illustrate, a researcher speculating on the effect of fertilisers on inner reef corals can modify the ontologies with assertions that "any reef in the vicinity of an agricultural coastal population will recover more slowly to bleaching or disease". The concept of an "agricultural population" could be precisely defined as a region with a low population density, or a region at particular set of coordinates, or a community with high sales in fertiliser products.

These statements are axioms that give the information required for a computer to logically infer results. Axioms are the truisms applied by the reasoning engine to automate the classification process. Therefore, some axioms can be suppositions or pure stereotypical speculations of the researcher. The version of "the world", or concept, depends on the hypothesis posed and is defined to the computer by the researcher. When the computer can "understand" a concept, that concept then can be used to automatically process the data to infer new knowledge.

The axioms may or may not be true of a real world but would be the monotonic suppositions in a specific hypothesis of which the suppositions would be stipulated in the research methodologies and assumptions. The hypothesis may also be more narrowly defined. For example, the researcher may unambiguously state the axioms: all towns which use fertiliser (brand "A") are agricultural towns and all towns with a population density more than 500 people/km^2 are non-agricultural towns. Then, when population density and fertiliser sales information are added to the KB as descriptive properties of a location, the reasoning engine automatically classifies a location to also belong to its equivalent class. More explicitly, any town that has sales in brand "A"

fertiliser and a small population density value will be subsumed to an "agricultural town class" and any town with higher population density will be subsumed to a "non-agricultural town class".

If a proposition requires conjecture on the part of a researcher the axioms in the KB can be changed to depict the proposed environment. Specifically, the researcher may wish to change the axioms of the KB to reflect a new view of "the world" which has changed from a previous conceptual view following a new line of enquiry or in light of new information. The modifications or additions are easily accomplished due to the modular design of the ontologies (refer to Chapter 3).

5.2.2. Data Integration and the OWA

The support for data integration is another benefit of semantic technologies. The requirement to more easily integrate data from disparate sources drives the development of the Semantic Web technologies. The data generally originates from research institutions, governments, non-profit organisations and commercial companies and is commonly stored in unconnected data repositories. Ontology-based data integration can be employed to bridge these data silo's (Wache et al. 2001). The ability to describe concepts and add context to data in a form that is decidable by the computer allows it to make the automatic links between concepts and the ultimate integration of the data. The disparate data can be automatically processed by the computer because the well-defined descriptions can add enough contextual information to data so that meaning can be inferred. Hence, the computer can make "intelligent" decisions or automate classification to link latent connections in the data, independent of the origin of the data.

Also, new information can be added to the KB based on changes in the researcher's line of query or as new data or information evolves because of the OWA. The addition of new information and unstructured data to a KB is expected under the OWA because the system assumes it never has a complete view of its world and there are always unknown facts to be added (Horrocks et al. 2003). As detailed in Chapter 2 §2.4.1.5.1, the OWA allows the KB structure to have an organic flexibility which can easily modify or adapt to new or additional concepts. In contrast to the OWA, a CWA assumes there is no other view of the world and supports negation as failure; that is, if information is not found within its structured data then it does not exist. Relational database systems are highly structured KR paradigms that uphold a CWA and so the addition of new fields to the schematic is a non-trivial task. The unstructured nature of the OWA offers the flexibility to make data integration a simpler process and is supported in logic systems employed by semantic technologies such as DL.

5.2.3. Inference Versus Query

Semantic technologies offer query functionality and also extensive inference capabilities that enable the automatic linking of data and make intelligent querying possible. A prime requirement for any data repository is the ability to query the data. Query capability is possible in semantic-based systems at either, or both, the RDF or OWL levels (O'Connor et al. 2007; McGuinness 2004). Currently, these levels require different query paradigms, SPARQL is used to query RDF triplestores (refer Chapter 2 §2.4.1) and SQWRL is used to query at the OWL DL level (refer Chapter 4 §4.3.2). Both semantic query levels can be applied in the Semantic Reef system.

There are two ways a Semantic Web model can answer questions: one is through queries the other is through inference. Inferred knowledge is derived from explicitly asserted facts as opposed to extracting knowledge via look-up or keyword search. Contextual meaning is added to the content contained by the computer through semantically modelled information. The added context can be processed by the machine to derive knowledge instead of text and can obtain more meaningful results through processes similar to human deductive reasoning and inference. Thus, computers can reason and automate information gathering to extract both explicit and implicit results (Brachman and Levesque 2004; Allemang and Hendler 2008).

Because the standard query paradigm relies on keyword search, implied relationships cannot be inferred. To infer these relationships reasoning capabilities are required. Semantic technologies support these reasoning functions and thus offer advantages to data processing and analysis. Conclusions can be inferred automatically through the expressivity afforded by the logic systems such as FOL and propositional logics. These systems use axioms to describe context about relationships clearly and unambiguously. The well-defined logical axioms are information the computer can employ to connect links that are not explicitly asserted. Therefore, the automatic linking of latent relationships with DL has potential benefits to automate processes that are currently accomplished manually in the data analysis stages of research.

5.2.4. Semantic Modularity

The constructs in semantic technologies have a component architectural nature that enables modularity and reusability, which is advantageous to flexible design. The hierarchical modular design of the Semantic Reef system is an example of component architecture that makes repopulation and reuse of the KB possible.

The reusable component of the Semantic Reef KB was developed initially as a set of atomic modules written in ontological form. The individual modules were composed and imported to a single holistic domain ontology (refer Chapter 3). The modular design adds far greater flexible and scalable functionality in the reuse of the ontologies. The separate ontology modules are imported to the higher domain ontologies and at each layer, higher complexity and finer granularity of domain concepts were added (Figure 3.2). The ontology hierarchy then became independent of any particular coral reef and its environment or human influential factors. In fact, as explained in Chapter 3, the system describes at a coarse level any coral reef in the world and can be reused by repopulating it with data pertaining to a specific reef. The reuse of the KB by repopulation, for example expunging the historic data used from the validation process and importing current data on any reef system, also effectively reduces the number of RDF triples. That is, because only data relevant to the hypothesis and specific location is imported the quantity of triples is reduced. This was a specific design parameter given the known performance problems of very large RDF data stores. Also, the flexibility is only possible due to this modular design of the KB.

Modularity also enables a more adaptable line of enquiry (Figure 5.1). To illustrate, coral bleaching was the theme in the validation process explained earlier, with a specific focus on a small sample of coral reefs within the GBR. If the coral reefs were different from those utilized in the validation, to infer a bleach-alert would simply require the repopulation of the KB (if the data were available). The "usable" domain-specific ontology at the higher layers of the KB hierarchy would be the only module greatly affected because the instance environmental data (SST, PAR, etc.) and geospatial and composition information (longitude, latitude, community composition, etc.) would be different. The inference rules would remain generally the same, with possibly small variations depending on climatology aspects and the lower "reusable" ontology modules would be unaffected.

Alternatively, and in contrast to the above example, if the line of enquiry is not coral bleaching, modifications would only be required at the higher usable layers of the KB. If the theme was not bleaching but instead regeneration rates, coral spawning or water quality, the underlying reusable ontology modules would still remain the same but the instance data and hypothesis rules would differ. To elaborate, all ontologies in the KB under and including the "Coral Reef" ontology are reusable modules and all above are usable domain-specific and application ontologies. A domain-specific ontology would be created for the new line of enquiry and it would import the lower ontologies. The KB is then repopulated with relevant domain-specific data. Then, at the highest level a separate application ontology, which contains the new proposals as inference rules, imports the domain-specific ontology and so hypotheses can then be posed.

5.3. Hypotheses Demonstrations

Exemplars are presented in this section to illustrate data integration, flexibility in hypothesis design, and the benefits of automated classification. To demonstrate the use of semantic methods and inference in hypothesis-driven research, the examples are intentionally simple. The purpose was not to prove or disprove an actual hypothesis but instead to use example hypotheses to illustrate the potential advantages and applications of the Semantic Reef system.

5.3.1. SST Indices with Live Data Flows

5.3.1.1. Methodology and Data

The validation stage tested the accuracy of the model by hind-casting the mass bleaching events of 1998 and 2002. To expand on this and portray the real time prediction potential of the system, the KB was primed with current SST data, streamed directly via the Web, and used to infer a coral bleach warning. The data were extracted for the 2008/2009 summer from three reefs. The reefs were chosen for this example as representative of the shelf locations: Cleveland Bay is an inner shelf reef system, Davies Reef is mid shelf and Myrmidon Reef is an outer shelf reef and all are in the central transect of the GBR.

The near real-time SST data in this exercise was streamed from the Davies Reef, Cleveland Bay and Myrmidon Reef monitoring sites (Kininmonth et al. 2004). The data was made available from the weather observing system, available through the AIMS data access portal and is a product of the AIMS data centre[33]. For this purpose, live data services were accessed by the Kepler workflow engine to populate the KR system for ontology-based data integration.

Each step of the workflow consisted of Kepler "actors" that provided access to the distributed data repositories and workflow libraries. The actors were directed to populate the KB with instances that have explicit date/time and temperature data-type property values. The workflow streamed SST data, both daily average and daily maximum, from the AIMS data centre, tagged it with a URI and then mapped the data to the domain-specific "GBR" ontology (Figure 5.2). The workflow actors also computed operations, for instance the LMST, the LMSM and MMM climatology values that are required for the inference rules were determined for each location. Separate temporal reef instances were created of each reef for each timestamp in the data

[33] http://data.aims.gov.au/

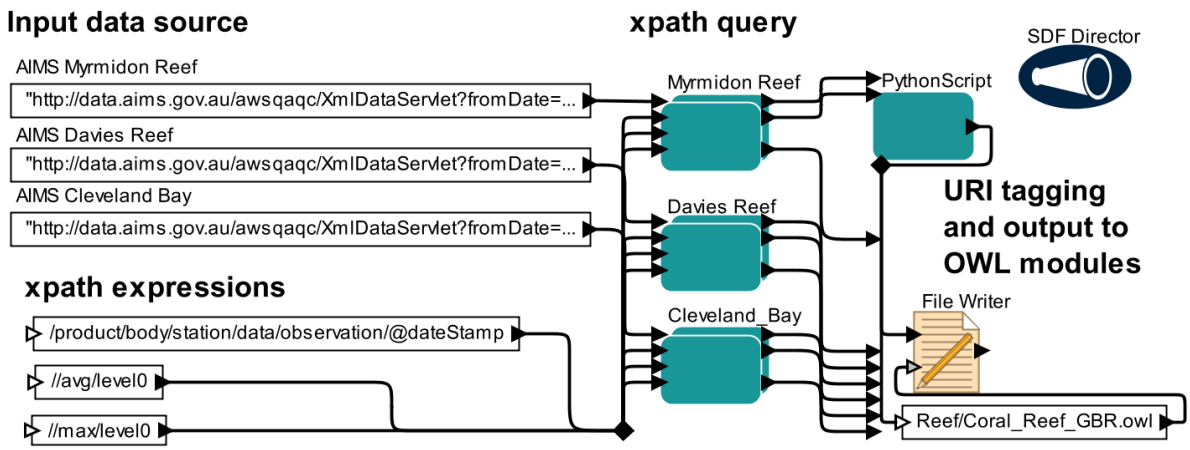


Figure 5.2 – A Kepler workflow for streaming SST data from AIMS, transforming remotely sensed data with XPATH actors to populating the KB.

mapping process and the relevant values were asserted to the temperature and climatology properties.

5.3.1.2. Results – Predicting a Bleaching Event

Based on the coral bleaching metrics: SST+, $_{Max}$SST, HotSpots and DHDs, the inference rules were applied to detect a problematic area. The rules automatically inferred any instances of a particular reef to belong to a categorised bleach watch class. Specifically, the Jess inference engine was invoked to apply the rules that imply a bleaching alert and instances that fit the bleach risk categories were correctly inferred to their respective classes (Figure 5.3).

A bleach watch outcome for Davies Reef in February of 2009 was the result of the inference rules. The graphs produced by the AIMS data centre[34] for the three reef locations during the 2008/2009 summer time period, which depict the average and maximum SST, were overlayed with the results from the inference rules (Appendix H). Cleveland Bay experienced moderately high temperatures until January but remained below the LMST threshold for the remainder of the summer. There were a number of warmer times during the summer period for Davies and Myrmidon Reefs, particularly Davies Reef. The NOAA Coral Reef Watch's Satellite Bleaching Alert (SBA) system (Figure 5.4) issued a bleaching watch alert on the 16 February 2009 for Davies Reef, which coincided with the inferred bleach risk instances from the Semantic Reef system.

[34] http://coralreefwatch.noaa.gov/satellite/sba_summaries/hist_davi.txt

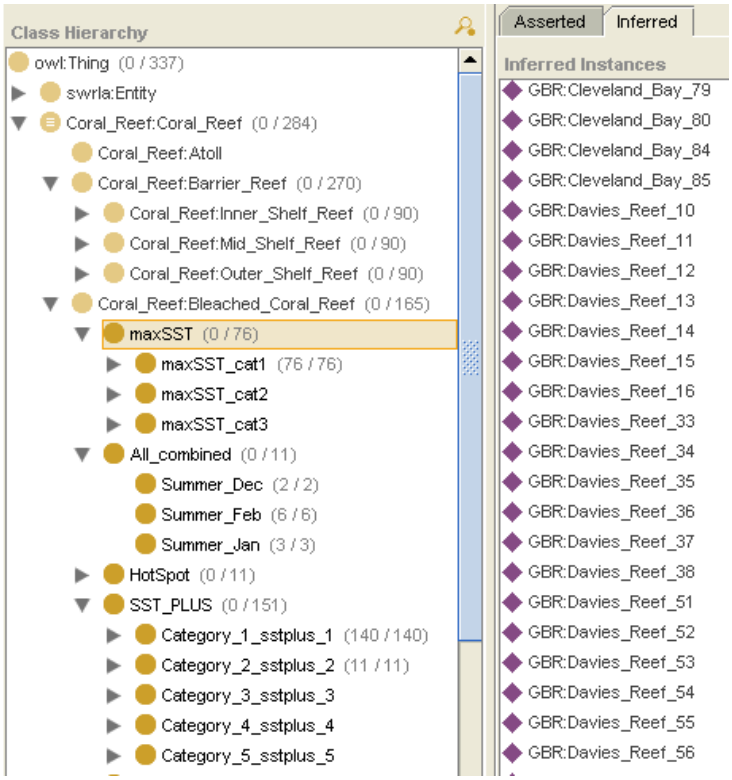


Figure 5.3 – The 2009 summer with SST data streamed from AIMS. The inferred results – instances are inferred to the correct Bleach Risk categories in the KB.

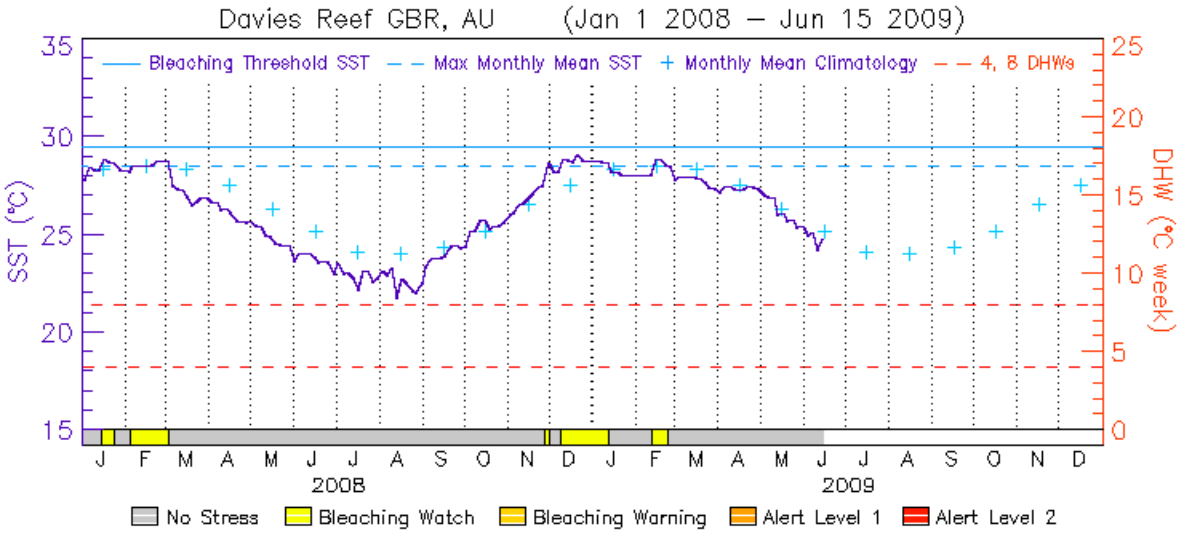


Figure 5.4 – The semantically inferred results (Appendix H) coincided with the 2009 bleach risk timeslots from the NOAA coral reef watch product, shown here for Davies Reef. A bleach watch was issued on the 16th of February 2009.

Davies Reef was surveyed recently (3rd April 2009) and there were no visible signs of bleaching. There are no recent survey reports available for Myrmidon Reef and Cleveland Bay because at the time of writing the *in situ* observations had not yet been conducted (AIMS 2009). However, it is assumed there was no bleaching occurrence because the temperatures for the summer period were not sufficiently extreme to invoke bleach warnings, via the standard measures (CSIRO 2007; NOAA 2009a) nor through the inference rules.

5.3.2. Applying Disparate Data to Theorise the Coral Bleaching Tipping-Point

To demonstrate ontology-based data integration, disparate data from a number of independent sources was mapped to the KB for inclusion in a sample hypothesis. The organisations that managed the data sources in this example include AIMS, NOAA's ICON/CREWS, the Australian Bureau of Meteorology[35] (BOM) and the Australian Bureau of Statistics[36] (ABS). Here, the ability to infer knowledge and find correlations in the data was shown through the combination of disparate data, workflow technology and SWRL inference rules.

5.3.2.1. Background

Bleaching is not uniform, but instead occurs in discreet reefs across the many reefs that comprise the GBR and other reef systems around the world. At present there is still only a limited understanding of the causal factors of bleaching, although sea temperatures are clearly involved. However, studies have shown it can also be caused by other factors such as salinity, light intensity, acidity, sedimentation, or even a combination of these factors (Brown 1997; Jones et al. 1998).

The coral bleaching predictions for the recent summer of 2008 and 2009 were mentioned in the previous section. The alerts, based on SST indices, were issued for the February period for Davies Reef from the NOAA Coral Reef Watch's SBA system (Figure 5.4). Although the high temperatures at Davies Reef invoked a bleaching alert, no bleaching occurrences were observed for the 2009 summer period (AIMS 2009). Instead, low levels of physical damage from storm action such as broken branching corals and overturned tabulate corals were observed. The monsoonal activity that enveloped the Queensland coastline for the greater part of February and March was one possible explanation for the erratic summer temperatures. The major contributing events were the category 1 tropical cyclone "Ellie" (30th January to 4th February) and the category 5 tropical cyclone

[35] http://www.bom.gov.au
[36] http://www.abs.gov.au/AUSSTATS

"Hamish" (4th March to 11th March) (BOM 2008). The reduction in temperature for the overall summer in the region was due to the low cloud cover and rain brought by these two events. However, despite the reprieve in high temperatures from the monsoons, the storms in January and February caused record levels of rain and extreme flooding. In fact, the bleaching observed at the inner reef systems was probably due to fresh water inundation during the 2008/2009 wet season (AIMS 2009). The compound effects of massive flooding flumes, heavy rainfall, harsh winds and earlier high temperatures on reefs between the Cairns to Whitsunday region have been thought to contribute to the bleaching that did occur (GBRMPA 2009).

Questions about coral bleaching are being posed to find the tipping point that leads to coral death. The questions entail the cumulative combination of ecological factors and stressors that contribute to the tipping point from a healthy coral to a dead coral via bleaching. The Semantic Reef system is a tool to pose such hypotheses and automate inferences of the available data and, therefore, is an appropriate method to theorise about the cumulative factors of bleaching. Once phenomena in the data are disclosed, *in situ* observations can be performed to confirm or negate the theory.

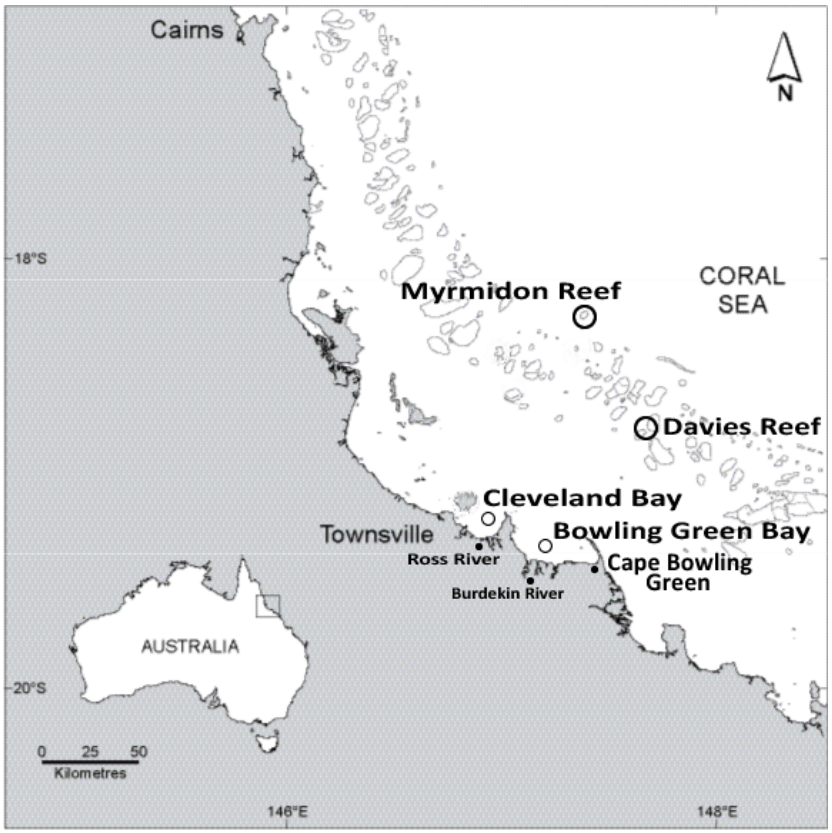


Figure 5.5 – The Townsville transect and the location of the reefs assessed in the demonstrations.

A selection of environmental and anthropogenic information was mapped to the KB to illustrate the data integration capabilities and flexibility in hypothesis design. The three reefs from the previous example (§5.3.1) are used here with the addition of a second inner reef system, Bowling Green Bay, which is monitored from the Cape Bowling Green monitoring station (Figure 5.5). The choice of reef systems depended on the public availability of the data.

5.3.2.2. The Environment Factors

The environmental factors incorporated in this test consist of average and maximum SST, Photosynthetically Active Radiation (PAR), Chlorophyll concentration and rainfall. The variety of data was derived from different origins for mapping to the KB. SST and PAR were extracted from the AIMS data centre for the four sites. The NOAA ICON/CREWS site provided rainfall and PAR data for Myrmidon and Davies Reef and rainfall data alone for Cape Bowling Green.

Representational data is common practice in marine research. For example, PAR designates the spectral range of solar light from 400 to 700 nanometres which is part of the photosynthesis process of plant life. The PAR information is a common proxy value for Chlorophyll-a, which measures the abundance of phytoplankton food sources. Another common representational proxy is rainfall data which is the proxy for salinity level approximation. Where appropriate, the gaps in the data were supplemented with representative data from a proxy location. For example, Townsville rainfall data was drawn from the BOM web site and was used as representative data for Cape Cleveland because data from the AIMS data centre was not available for this time period.

5.3.2.3. The Anthropogenic Factors

An anthropogenic influence may be of interest or significance in an arbitrary hypothesis. The anthropogenic factors in this test were human population and were included to exemplify the diversity of information that can be introduced to the system for hypothesis testing. Thus, data about the population density and quantity, for the coastal transect from Townsville to the lower Burdekin, was included.

The population data for Townsville, Thuringowa and the Burdekin was extracted from the ABS online database and downloaded to the workflow. The data included the geographic figures and demographic breakdowns (age and gender). Questions to theorise about the effects on coral reefs as a result of the human coastal population density may be posed. For instance, one may question the make-up of a specific demographic group and how they affect the local coral reefs. A research question could be “how does a particular group’s use of sewage influence coral health?”.

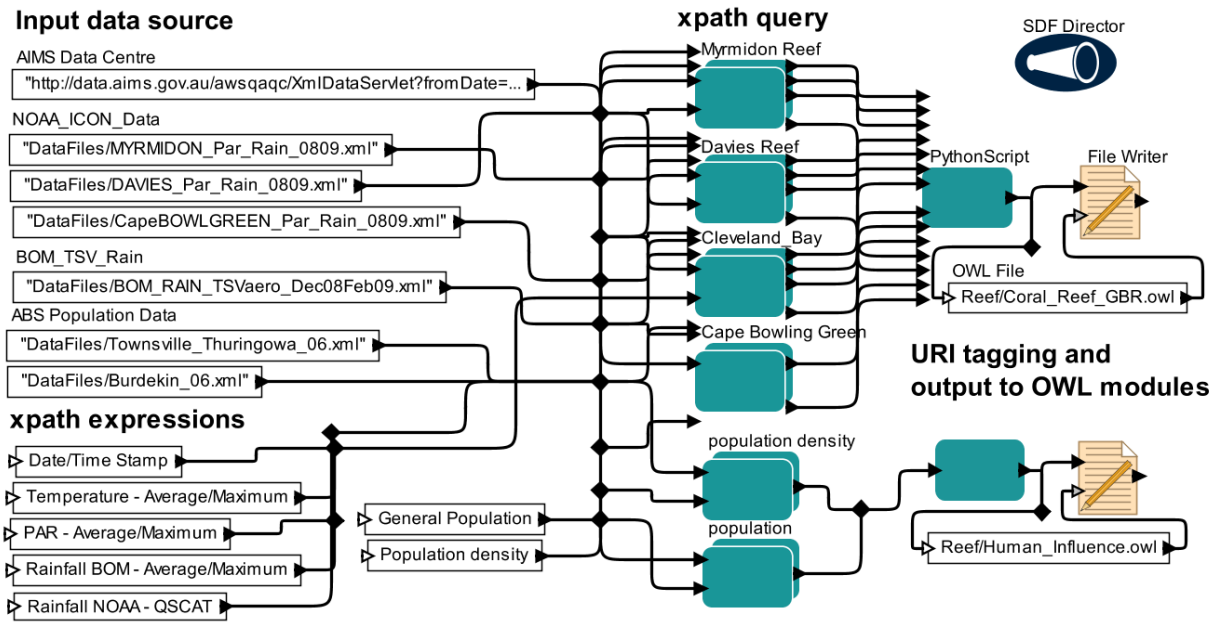


Figure 5.6 – A Kepler workflow to populate the KB with PAR, rain, salinity and SST data from AIMS, NOAA and BOM and human population quantity and density from the ABS.

The Townsville/Thuringowa and the Burdekin locations were appropriate for this test because they both have river outlets: the Ross River and the Burdekin River, respectively. Because the quality characteristics of these local rivers differ, hypotheses can be demonstrated that examine water quality and coral health. Data on the two inner shelf reefs, Cleveland Bay and Bowling Green Bay, were included because they are the outlet locations for two major river systems in North Queensland (Figure 5.5). The Burdekin region is predominantly agricultural with a low population density and the Burdekin River opens onto Cape Bowling Green. In contrast, the Townsville region is industrial with a high population density and the Ross River opens into Cleveland Bay. The AIMS Automatic Weather Stations (AWS) at Cape Bowling Green and Cleveland Bay monitor the respective river mouths.

5.3.2.4. The Workflow – Data, Methodology and Assumptions

The Kepler workflow imports and transforms the disparate data and prepares the KB by populating the ontologies (Figure 5.6). The data from the four disparate data sources is manipulated via the Kepler XPATH and Python actors. The XPATH expressions and queries extract the specific data values from the data streams. The data is then converted to an array of values and sent to a Python[37] scripting actor. These actors then implemented simple scripts that

[37] Python is a general-purpose high-level programming language: http://www.python.org/.

were written to tag each value with a unique URI. The data was then ready to send to the KB to populate the appropriate ontology modules (Figure 5.6).

Location	***SST (average and maximum per day)***	***PAR (average per day)***	***Rainfall Data (total per day)***	***Human Population***
Cleveland Bay	AIMS	AIMS (Cape Bowling Green)	BOM	ABS (Townsville/ Thuringowa)
Cape Bowling Green	NOAA	AIMS and NOAA	BOM (Cleveland Bay via Townsville Aerodrome)	ABS (Burdekin)
Davies Reef	AIMS	AIMS and NOAA	NOAA	
Myrmidon Reef	AIMS	AIMS and NOAA	NOAA	

Table 5.1 – Matrix of the available data sources as retrieved and distributed by the workflow.

The KB was populated with data from four disparate data sources at this stage and is depicted in Table 5.1. The details of the data were as follows:

- No SST data were available from the AIMS data centre for Cape Bowling Green because it is a land-based data station. Therefore, the SST satellite data available from NOAA ICON represents the SST for Bowling Green Bay;
- The PAR figures for Cleveland Bay from the AIMS data centre read as 0.0 or 0.1 for the entire summer months and were assumed to be inaccurate. Therefore, the PAR data available from AIMS for Cape Bowling Green was substituted as representative PAR information for Cleveland Bay;
- No rainfall or salinity data were available from the AIMS data centre for any of the four locations;
- PAR and rainfall data were available for Davies Reef, Myrmidon Reef and Cleveland Bay from NOAA ICON;
- Only PAR data were available for Cape Bowling Green from NOAA ICON;
- The rainfall data for Cape Bowling Green were representative data available from the BOM via the Townsville aerodrome weather station; and
- The population data for the ABS were from the 2006 Census and were assumed to be representative of the current status.

5.3.2.5. The Logic and Rules

The questions posed of the current KB are now quite flexible because the researcher using the system does not need to have a predetermined hypothesis. This flexibility is the focus of this illustration and, hence, the examples presented as observational hypotheses are purposefully simplified. The propositions that could be posed would infer any instance in the KB to be disclosed as anomalous, depending on the environmental data values. The KB currently holds PAR, rainfall and SST information on four reefs and also the human density of the coastal regions that are in proximity to two of these reefs. An example rule to indicate the level of PAR and accumulated rainfall, in correlation with SST, as a potential mix of causal factors responsible for a coral bleaching, is written in SWRL as:

```
Coral_Reef:Coral_Reef(?x)  ∧
Coral_Reef:hasLightEinsteinsOf(?x, ?par)  ∧
swrlb:greaterThanOrEqual(?par, 500) ∧ swrlb:lessThan(?par,750)
∧ Coral_Reef:hasWeeklyRain_mm_Of(?x, ?rain)  ∧
swrlb:greaterThanOrEqual(?rain, 20)  ∧
Coral_Reef:hasDailyAverageSSTof(?x, ?meanTemp)  ∧
Coral_Reef:hasAverageLongTermSeaSurfaceTemperatureOf(?x,?LMST)
∧  swrlb:greaterThanOrEqual(?meanTemp, ?LMST)  ∧
Reef_Stock:Coral(?partCoral)  ∧
Coral_Reef:hasPart(?x,?partCoral)  ∧
Trophic:is_Hermatypic(?partCoral, true)  ∧
Trophic:hasGrowth(?partCoral, Trophic:fast)
→ Coral_Reef:Observe_Reef(?x)
```

The rules were fashioned as observational hypotheses; therefore if any phenomena in the data were uncovered the location could be observed for *in situ* confirmation of the hypothesis. If there were a change in the hypothesis due to new information or an epiphany, the rules could be modified to express the new hypothesis simply by adding or removing antecedents to the rules. For example, the human population affects the reef systems and in particular the inner shelf reefs. The human influence can be a measure of population density in the coastal regions and, given data on other factors, the extremity and ramifications of the influence could be hypothesised as a contributing factor to coral bleaching. An inference rule could question and extract the variations in human populace in correlation with other prescribed factors as exemplified in the second SWRL rule:

```
Coral_Reef:Coral_Reef(?x)  ∧
Coral_Reef:has_Human_Influence(?x, ?y)  ∧
Human_Influence:Influence(?y)  ∧
Human_Influence:hasPopulationDensity(?y, ?pop)  ∧
```

```
swrlb:greaterThan(?pop, 500)  ∧
Coral_Reef:hasLightEinsteinsOf(?x, ?par)  ∧
swrlb:greaterThanOrEqual(?par, 500)∧ swrlb:lessThan(?par,750)
∧  Coral_Reef:hasDailyAverageSSTof(?x, ?meanTemp)  ∧
Coral_Reef:hasAverageLongTermSeaSurfaceTemperatureOf(?x,?LMST)
∧  swrlb:greaterThanOrEqual(?meanTemp, ?LMST)  ∧
Reef_Stock:Coral(?partCoral)  ∧
Coral_Reef:hasPart(?x, ?partCoral)  ∧
Trophic:hasGrowth(?partCoral, Trophic:fast )
→ Coral_Reef:Observe_Reef(?x)
```

The inferred instances of this rule were locations in proximity to a medium to high human coastal population, which in this case is the Townsville region. The rule sets PAR in the higher wavelengths, SST higher than average and a high rainfall, which is a proxy for a low salinity percentile. Consequently, the temporal reef instances that fit this combination of influence and environmental values could be observed in situ for signs of bleaching.

5.3.2.6. Results

The KB was populated by the workflow with 360 instances, 90 for each of the four reefs. One temporal instance was created for each day per reef in the 2008/2009 summer period and linked to the population quantity and density human influences for the location.

The first example rule extracted any instance that had a PAR over 500nm, a category one or above SST+ and a high rain fall, simultaneously. Two instances were disclosed: Cleveland Bay and Cape Bowling Green on the 1st of December. Because both locations share the same representative PAR and rainfall data it was expected both inner reef systems would have similar results. Small changes to the hypothesis and subsequently its inference rule can be easily made and result in changes to the outcomes for observation. To illustrate, the rainfall antecedent of the first rule was lowered from 20 mm to 10 mm. The outcome was changed and disclosed Myrmidon Reef on the 16th of December to also be an instance of interest for observation. Further, when the parameters of the PAR factor were lowered from 500 to 300, this time leaving the other factors (rainfall and SST) the same, the inference results disclosed instances of Davies Reef for 23rd and 3rd of February for observation.

The second rule focused on reefs in proximity to high density population. The KB contains instances of only two locations which contain inner shelf reef types (i.e., Cleveland Bay and Cape Bowling Green). These areas have been connected to the human influence type (population) through the assertion of the object property “has human influence”. Therefore it was expected that the conclusion from the second rule would infer only instances in the highly populated location

(i.e., Cleveland Bay area) to the "Observation" class. The results from the rule exposed instances from the 1st to the 16th and the 21st to the 28th of December 2008 at Cleveland Bay, which aligned to the December bleaching signs.

Due to the gaps in the available data, the scope of the example hypotheses and the demonstration of versatility in hypothesis design and data integration, shown here, were constrained. The questions required a mix of environmental factors that may induce bleaching and, hence, the gaps in the data were supplemented with representative data, which might not produce correct outcomes. For instance, it PAR data for Cleveland Bay was represented by PAR from Cape Bowling Green and BOM rain data from Townsville aerodrome (3.5km from shore) and it was assumed that this was representative of Cape Bowling Green rainfall. The methodology and assumptions of a specific hypothesis would need to include these specifications on the validity of the representational data in the research documentation.

If data had been available for all four sites of each environmental variable, including the obscure factors such as CO_2 concentrations, the outcome of the hypotheses tested here may have been of great interest to the coral bleaching community and not simply a demonstration of the system's capabilities.

5.3.3. Classifying the GBR – by Community Makeup and Location

5.3.3.1. Background

Inferred knowledge is derived from well-defined asserted facts that describe a concept. The semantically modelled information adds context to data and through reasoning and inference it is possible to obtain more meaningful results. The facts may, or may not, be directly linked to information in the KB. However, upon reasoning over the data, both explicit and implicit linkages can be obtained. (Brachman and Levesque 2004; Allemang and Hendler 2008).

Current marine research methods typically account for data gaps by the use of data from one sensor to be representative data for other surrounding reefs or for other reefs that are similar by reef type. To use representative data for reefs that surround a monitoring station is necessary because it is often unfeasible to have a sensor station on every reef due to the cost of remote monitoring and the ecological interruptions involved. The data can represent reefs that are similar by their type and not just their location. Some example models to describe or represent a reef type include: by the community make-up, by thermal sensitivity and/or by nutrient levels (Maynard,

Anthony et al. 2008). Hence, based on characteristic models rather than only proximity, data from one reef can be indicative of others via this method.

Automatic reasoning and classification is applied in the following demonstration to show how these models may be integrated into the system. Logical axioms are explicitly expressed to describe the reef types by the thermal sensitivity of the community composition and by location. Then, once the reasoner had automatically linked the implicit connections, all instances of the KB were automatically subsumed to belong to numerous classes simultaneously.

5.3.3.2. Classifying Reef-Type by the Community Mix

The make-up of a reef-type by its community composition and sensitivity to heat stress can be described in the following statements[38]:

- Type A reef - is a reef with a high percentage of slow growing coral and is therefore thermally tolerant; whereas,
- Type B reef - is a reef with a high percentage of fast growing coral and is therefore thermally sensitive.

Unless specific reefs were manually asserted to belong to a Type A or B reef class, a query to select all sensitive reefs would not return any results. However, the thermal tolerance assumptions can be defined in an ontology as "necessary and sufficient" axioms of a Type A and Type B reef class. Specifically, to define the two reef types, axioms can be written as:

```
Class: ReefType_A Defined
       SubClassOf: Coral_Reef
       EquivalentTo: hasPart some (Coral
             and (isSlowGrowing hasValue true)
             and (percentageFastCoralCoverage hasValue < 30))
Class: ReefType_B Defined
       SubClassOf: Coral_Reef
       EquivalentTo: hasPart some (Coral
             and (isSlowGrowing hasValue false))
             and (percentageFastCoralCoverage hasValue > 60))
```

The object property "hasPart" is asserted to link coral reefs with, among others things, coral instances.

[38] Based on the AIMS Data Centre's "Future of the reef" model http://data.aims.gov.au/reefstate/sci/reefstate/

The Boolean data type property "isSlowGrowing" is set to true or false depending on the nature of the coral, for example:

```
Individual: Porites_cylindrica
Facts: isSlowGrowing hasValue true
       and isPartOf Davies_Reef
Individual: Acropora_formosa
Facts: isSlowGrowing hasValue false
       and isPartOf Myrmidon_Reef
```

The values to describe the coral instances were set and then introduced and asserted to a particular coral reef class. These property restriction axioms were added to a selection of reefs to illustrate the automated inference capability of the system. The selection included reefs from different types which in this case, were fringing and barrier reefs and from the different locations: inner, mid and outer shelf areas (Figure 5.7). The delegation of the community mix of fast growing and slower growing coral to each reef was arbitrarily appointed for the purpose of the demonstration.

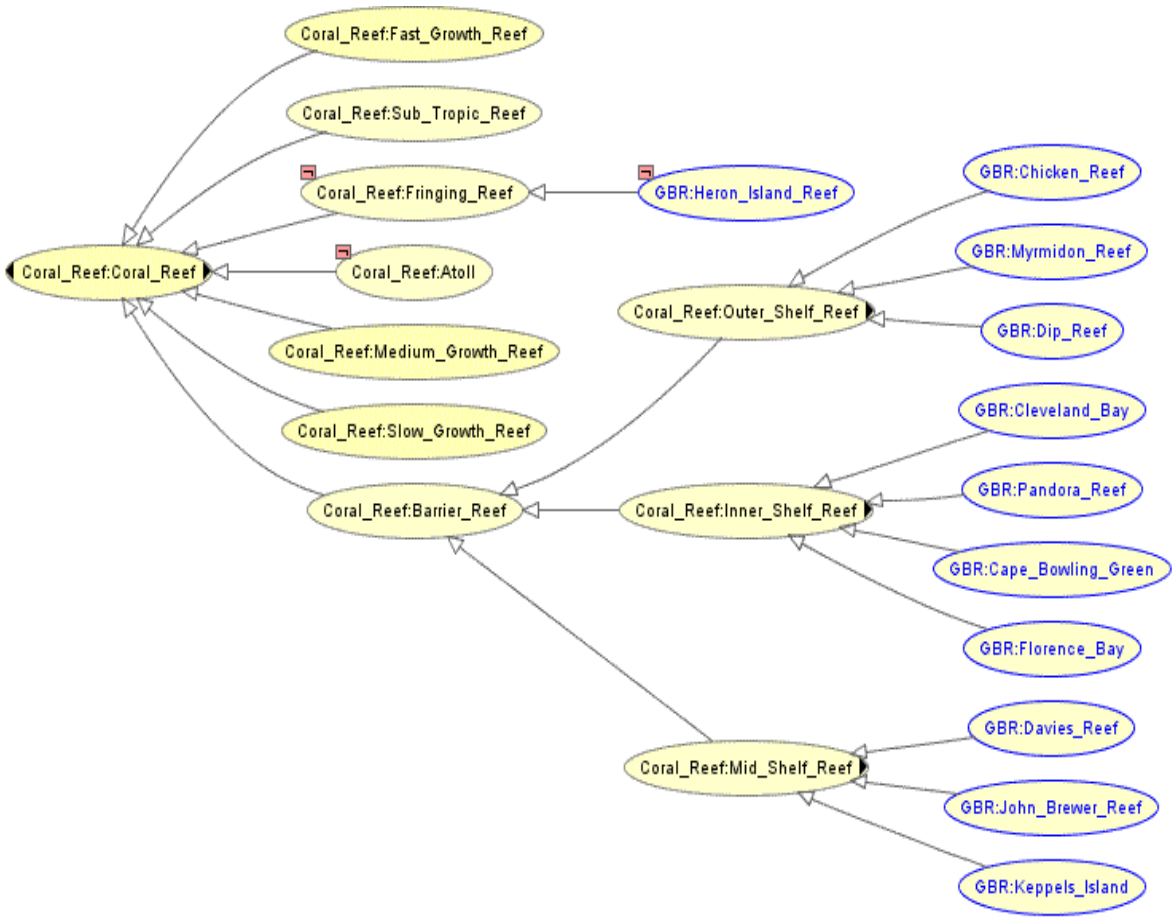


Figure 5.7 – A select segment to depict the classification before the Pellet reasoner. The reefs are designated as subclasses of major reef types (e.g., barrier, fringing, atoll, etc.).

When the reasoner finished classifying the ontologies, the reef classes were correctly inferred to the various reef-type classes. All reefs that have a greater percentage of fast growing coral were subsumed to automatically belong to the "Fast Growth Reef" class and the reefs that had been appointed with a higher mix of slow growing corals were subsumed to belong to the "Slow Growth Reef" class (Figure 5.8). Then, once all reef classes were subsumed to belong to a variety of functional types, inference rules were posed based on the data in correlation to the reefs characteristics as well as environmental factors.

5.3.3.3. Classifying Reef-Type by Location

There is a consensus that the location of a coral reef in proximity to other reefs has environmental commonalties (Sweatman et al. 2003). This methodology, which is predominately

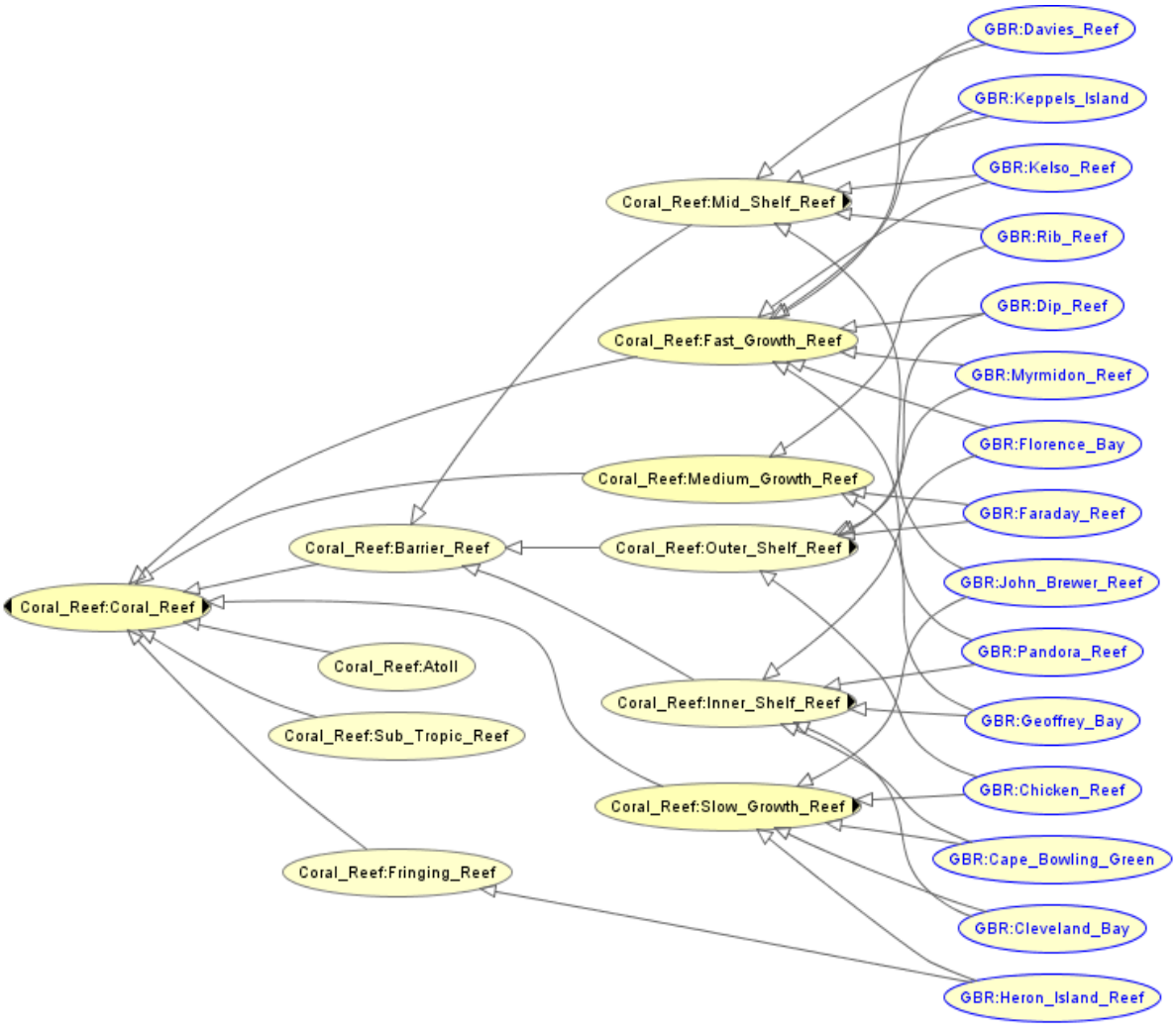


Figure 5.8 – After classification with the Pellet reasoner the reefs were subsumed to belong to the correct reef type (according to arbitrary axioms).

related to environmental factors (e.g., pH level, SST, PAR, etc.), assumes that the factors of a specific geospatial transect will be similar and indicative of each reef in that transect. There are many examples in literature of hypotheses posed of reef-types modelled by location (AIMS 2009; Maynard 2004). This example applied a similar technique to the previous example; however, here, the reef classes were inferred to belong to a reef-type class based on its specific geospatial grid location.

Reef-types by geospatial values were used to illustrate the systems' automated classification of reefs. GBRMPA designates gridded areas of the GBR for both management and research purposes. The GBRMPA grid[39] is based on longitude and latitude values and divided into a matrix from North to South and inner-shore to outer-shelf. To define each gridded area of the

Figure 5.9 - Reef types classified by the Pellet reasoner to belong to the respective "reef type" model – Grid location, a fast growth composition and shelf location.

[39] http://www.gbrmpa.gov.au/corp_site/management/zoning/zoning_maps.html

marine park, geospatial property values were declared for each coral reef class. The longitude and latitude property restriction axioms test the asserted values to subsume a coral reef to also belong to the grid regions.

Because the proximity boundaries are defined by longitude and latitude coordinates of a gridded region in a reef system, the reasoner relies heavily on the range reasoning function to classify the reef-type by location. Currently, the "GBR" domain ontology defines the reef-types by grid location as classes with asserted "necessary and sufficient" property restrictions about the "longitude" and "latitude" data-type properties (Figure 5.9). On reasoning over the KB the reefs of the GBR are subsumed to different reef-types which can effectively add other dimensions when questions are posed of the system.

The functional capabilities to define concepts are being extended as the emerging OWL standards are developed. Currently OWL 1.1 has no support for advanced data-type restrictions such as testing for a range of asserted data-type properties. However, OWL 2.0 has added this functionality and has been implemented in Protégé 4 (Protégé 2009). When automating the linkages between reefs by type, these new developments will add extra functionality for richer descriptions and flexibility in restriction definitions of data–type properties.

5.4. Discussion and Summary

This chapter expands on the validation of the Semantic Reef system by demonstrating the potential of semantic technologies to enable research through observational hypotheses. Features of the Semantic Reef system were demonstrated which included data integration capabilities, ontology reuse, flexible hypothesis design and the coupling of semantic inference with automated reasoning and classification.

Data from disparate sources, both live and static were mapped to the KB. The demonstrations began by repopulating the historic data from the validation test in Chapter 4 with live data illustrating the reuse of the KB. This experiment showed how live data could be mapped to the KB for real-time inference and predictions, given the availability and access to Web services and data. SST data was streamed directly from the AIMS data centre to populate the KB and inference rules that represent the bleaching indices were applied to infer a coral bleaching event. Because the inference rules mimic the current standardised metrics for predicting a bleach event the results were expected to be, and were, in-line with the current predictive tools from ReefTemp and NOAA's Bleach Watch.

The next example demonstrated data integration and flexibility in hypothesis design through a variety of rules applied to a combination of disparate data. Environmental data from AIMS, NOAA and the Australian BOM and anthropogenic data from the ABS was imported to the system. Once the unrelated data was coupled to the ecological and environmental ontologies, querying and propositional testing, through inference, was performed.

The flexibility afforded to the researcher to pose questions is only limited by the information in the KB at that time. The more disparate information introduced to the system the higher the likelihood of inferring unusual correlations in the data and, due to the OWA which assumes not all is known of the world, the addition of new concepts and data to the KB is trivial. For example, other factors can be added to the system such as scheduled dredging, water quality (e.g., toxicity of runoffs, CO_2 concentrations and pH, etc.) or more obscure factors (e.g., agricultural fertiliser sales). Questions can then be asked of the data to investigate the tipping point or the cumulative mix of causal factors that are the difference between healthy coral and corals killed by bleaching. The demonstration proved the system can support flexible hypothesis design and is able to disclose or extract anomalous instances from data. The anomalies can then be observed *in situ* to prove a hypothesis correct or incorrect.

Links between all concepts in a Semantic KB do not need to be manually asserted; the reasoning engine can make the latent connections automatically. The system's reasoning capability was demonstrated with explicit descriptions of two reef models that automatically classified a reef by its community composition and thermal tolerance or by its location and proximity. Once classification of the KB was complete questions could then be posed of a reef through inference by its ecological or geographical model.

To realise the longer-term objectives of the Semantic Reef project, where the architecture can be used in a wider range of hypothesis driven research and/or sensitivity analyses (using a range of different data sources), differing logic systems will be required. One possible avenue for investigation will be to incorporate Bayesian logics and to extend the sensitivity analysis capabilities with probabilistic logic. Currently, solutions to permit modelling for uncertainty are not in mature stages of development and at this stage remain future work.

The following chapter will assess the functionality of the Semantic Reef system through a performance analysis. The analysis included a series of performance tests that examine the functionality and scalability of the system relative to the quantity of instances and triples imported to the KB.

Chapter Six

The Architecture and the Quantifiable Test of Functionality

6.1. Chapter Synopsis

The previous chapter explored the differences semantic applications offer in data integration and exploration methods. Semantic inference was applied in hypothesis-driven research exemplars to trial the Semantic Reef system. The demonstrations confirmed the advantages available for data integration, flexibility in hypothesis design and the automation of processes, such as linking data and concepts for representational research methods. The prototype offers potential as a tool to hypothesise over disparate unconnected data and to determine gaps in current and future data.

This chapter explores the feasibility of the Semantic Reef system as a hypothesis tool to enable researchers in the coral reef domain. The reasoning and inference functionality of the system was tested to prove the viability of the system as a desktop tool that can be used to infer information or disclose phenomena in data.

The performance analysis methodology includes descriptions and justifications of the software and hardware computing platforms used for the tests. The relevant specifications of the software and hardware tools chosen for incorporation in the Semantic Reef architecture are evaluated and justified. The performance analysis consisted of a series of tests administered in a simulated computing environment indicative of a researcher's *in silico* environment. The test scenarios focused on the quantity of data or triples introduced to the system versus the time to load the KB and then reason and infer over the KB.

The results and discussion detail the outcome of the performance analyses. The viability of the Semantic Reef system as a hypothesis tool showed quantitatively positive within the limitations of the computing environment on which it is deployed.

6.2. The Performance Analysis Methodology

In this chapter the term platform refers to both the software and hardware combined. A hardware platform consists of the processing, storage and memory combination and capacity of a computer and the software platform consists of the programs that run on or operate the hardware. The desktop computing platform is common in computing environments such as the client/server networks found in organisational computing deployments.

The Semantic Reef system is a hypothesis tool intended to be deployed on a desktop computer initially. Marine researchers typically use standard desktop computers to run *in silico* analyses that do not require massive processing power. Therefore, the methodology for testing the functionality of the system had to simulate the probable environment of researchers as they explore possible and probable causes for phenomena.

6.2.1. The Computing Platform for the Performance Analysis

The specifications relevant to the performance analysis are shown in Table 6.1 and include the hardware platform and the software tools applicable to the Semantic Reef architecture. The performance analysis required a consistent hardware platform to eliminate bias when comparing the variables and results, so the parameters such as processing power, memory, and software platform remained constant. To maintain a constant value for these testing parameters the same test platform was used and it simulated a researcher's computing environment. The trial computer was a standard desktop machine running a Microsoft Windows client operating system (Table 6.1).

Test Platform – Hardware and Software			
Operating System	***RAM***	***CPU***	***Java Runtime Environment***
Windows XP PRO 32-bit	2 Gigabytes	Intel Core 2 Duo 1.86 GHz	Version 1.5.0_12-b04
Software Tools			
Feature	***Name***	***Version***	
Ontology Editor	Protégé	3.4 (build 130)	
		4 (build RCI)	
Reasoning Engines	Pellet	1.5.1	
	FaCT++	1.3	
	RacerPRO	1.9.2	
Inference Engine	Jess	71p2	
Scientific Workflow	Kepler	1.0.0	

Table 6.1 – Specifications of the computing platform and the software tools incorporated in the performance analysis of the Semantic Reef architecture.

6.2.2. The Knowledge Base Software

Protégé is a free open source ontology editor and KB framework (Protégé 2009). The Semantic Reef architecture consists of the combination of scientific workflow tools and a KB infrastructure. The Protégé framework was chosen as the basis of the KB infrastructure because it offers wide developer and user support through an active development community for use-case applications such as the Semantic Reef. The justifications for choosing Protégé as the ontology and KB development tool within the Semantic Reef architecture are as follows:

- Due to the open source and the Mozilla Public License (MPL), Protégé is free to use, as opposed to other KB frameworks and ontology editing tools which are proprietary software.
- A high degree of support is offered by a strong community via active discussion forums for developers, professionals and students.
- Development via the Application Program Interface (API) is possible because the open source nature of the licence permits full access to the source files.
- Protégé is based on Java and provides a flexible development base for "plug and play" environments. In fact, Protégé's plug-in architecture can be extended through the Java-based API for building knowledge-based tools and applications.
- Protégé offers a range of direct and indirect support for a variety of reasoning engines, such as Pellet (2007), FaCT++ (2008) and RacerPRO(2008).
- Protégé is being developed in synchronisation with the Semantic Web standards and recommendations (e.g., OWL, OWL 2.0, SWRL, etc.). In fact, key people involved in the development of the Protégé project are also seminal in the development of the standards.

The Protégé project currently has two framework versions available, Protégé 3.4 and Protégé 4, which are being developed concurrently. Table 6.2 shows a comparison of the components available in the separate versions that are relevant to the Semantic Reef system's development.

Function / Version	Reasoning Support			Inference Support			OWL	
	Pellet	FaCT++	RacerPRO	SWRL	SQWRL	Built-in Library	1.1	2.0
Protégé 3.4	Yes - direct in memory connection	Yes - via DIG 1.1 interface	Yes - via DIG 1.1 interface	Rule Engine Bridge to Jess	SWRL API	SWRL API	Yes	No
Protégé 4	Yes - direct in memory connection	Yes - direct in memory connection	No	Via Pellet (limited)	No	No	Yes	Yes

Table 6.2 – Matrix to compare specific components in Protégé 3.4 and Protégé 4 relevant to the Semantic Reef architectural development.

6.2.2.1. Protégé 3.4

The advantages and functions provided by the Protégé 3.4 platform and relevant to the Semantic Reef system are as follows:

- Protégé 3.4 offers a direct link to the Pellet reasoning engine and a Description Logic Implementation Group (DIG) version 1.1 interface link for compliant DIG reasoners (i.e., FaCT++ and RacerPRO).
- Protégé 3.4 supports SWRL inference rules using the Jess inference engine. The SWRL rule engine bridge is a component of protégé 3.4's SWRL implementation (i.e., the "SWRL Tab") that offers an integrated connection between an OWL model with SWRL rules and the Jess rule engine (Jess 2006; O'Connor et al. 2005).
- The SWRL Tab offers extra support with emerging components such as SQWRL queries (discussed in Chapter 4), which is a query language that extends SWRL to support querying of OWL ontologies (O'Connor et al. 2007; SQWRL 2008).
- The SWRL Tab supports *built-ins*, which are defined predicates that can be atoms in a rule. There are a number of core SWRL *built-ins* defined in the SWRL submission (2004) but users may also define their own libraries through the SWRL API's built-in bridge (O'Connor et al. 2008).
- The SWRL API also provides sets of built-in libraries that comprise developer's implementations of core SWRL built-ins. Some of these implementations include a temporal library for reasoning with temporal information, ABox and TBox libraries and a mathematical library for basic calculation (O'Connor et al. 2007; O'Connor et al. 2008).

6.2.2.2. Protégé 4

At the time of writing, some functions provided by the Protégé 4 platform differ from Protégé 3.4, and were important in deciding which framework to adopt in the development and testing of the Semantic Reef architecture:

- SWRL is supported via the Pellet reasoning engine; however the implementation is limited. No support is available for bridging to other inference engines (i.e., Jess) or support for built-ins, either core SWRL built-ins or user developed.
- No support was available for SWRL built-in libraries at the time of this writing; however there are discussions that the SWRL Tab will be integrated with the Protégé 4 framework in the future.
- Protégé 4 offers a direct link to the FaCT ++ and Pellet reasoning engines but does not support a DIG interface.
- Protégé 4 supports OWL 2.0, which is the new W3C recommendation for the OWL standard (Grau, Horrocks, Motik et al. 2008).

An important factor in selecting the KB framework was the reasoning support. There are two means to initialise the reasoning engine through Protégé; indirectly, through the DIG interface (Bechhofer et al. 2003) and directly through an inline memory connection. The DIG interface provides a communication connection to any DIG compliant reasoner (e.g., Pellet, FaCT++, RacerPRO, etc.), which is an advantage over the indirect method. However, the primary disadvantage of the indirect access is the lack of support DIG 1.1 has for data-type properties. Protégé 3.4 has reasoning support via the DIG interface and, in this case, RacerPRO was the reasoner chosen for the trials. Alternately, through the direct in-memory connection of the KB framework, the FaCT++ reasoner is available in Protégé 4 and the Pellet reasoning engine is available to both Protégé 3.4 and Protégé 4. Therefore, both versions offer adequate availability to reasoning support.

Notably, both Protégé 3.4 and Protégé 4 were implemented during the ontology development and the performance trials for a more complete comparison of reasoning outcomes. Because of the extra reasoning support in Protégé 4 (i.e., direct access to both Pellet and Fact++ reasoners), the system could be tested using three reasoning engines. The greater range available meant the results for the performance analysis tests were more extensive and non-biased.

Another significant function, only available in Protégé 4, was the degree of support for the new OWL standard, OWL 2.0. The new version of the OWL standard includes, among other components, more extensible support for data-type property manipulation. Because of the more extensive abilities that were added for reasoning with data-type properties, OWL 2.0 is more flexible in concept descriptions. Importantly, the modelling capabilities of OWL will be enhanced with the new aggregation and comparison functions and the ability to express ranges of values as restrictions on a class.

The comparison facility in OWL 2.0 is significant to the capabilities of the Semantic Reef system. The automatic classification of a reef to a "reef-type by location" requires the comparison functionality because if the reasoner is to classify and subsume a reef to a gridded location such as the longitude and latitude coordinates, it needs the ability to reason over a range of values. Data-types such as integers, floats and temporal values can be compared and inferences made with the new OWL 2.0 standard, based on ranges of these data-type values.

However, although reasoning with data-type properties is important, SWRL functionality was a crucial requirement. Protégé 3.4 was chosen as the main infrastructure for the Semantic Reef architecture due to the extensive support for SWRL inference. The need to implement and apply complex SWRL inference rules in Horn logics for inferring new knowledge or extracting phenomena was more important to the immediate goals of the Semantic Reef architecture. Notably, the Protégé 4 framework will be increasing its support of SWRL functionality in future developments. Then, in addition to OWL 2.0 support, Protégé 4 would be the appropriate infrastructure as the editor and KB framework for the Semantic Reef project.

The performance analysis centred on the scenarios from Chapter 5 and compared the quantity of data versus the processing time. The tests were run over the same time frame: the three month summer period from the 1st December to the 28th February (2008 – 2009) with datasets at daily and hourly intervals for up to four reefs. The tests were varied using a matrix of attributes, which could be changed by factor and the outcome then compared for each run of the system (Table 6.3). Also, by processing reasoning and inference functions over a growing range of triples, the limitations and performance of the desktop computing environment was tested.

Attributes	***Comparison Matrix***																				
Number of reefs	3										4										
Temporal intervals	Daily					Half hourly					Daily					Half hourly					
Number of property assertions	13		26			13		26			13		26			13		26			
Number of inference rule atoms	5	9	5	9	16	5	9	5	9	16	5	9	5	9	16	5	9	5	9	16	

Table 6.3 – A matrix of the testing attributes – the variations in the growth of triple and reef instance quantity.

The processing time factors for the experimental performance runs are as follows:

- The time taken to load all triples to the KB.
- The time taken to load the complete KB, including triples, lists, classes and properties.
- The time taken to reason over the KB with Pellet, FaCT++ and RacerPRO. This test was completed for two ontology levels, both at the "usable" level of the hierarchy: the domain-specific ontology ("GBR.owl") and the application ontology ("GBR_Rules.owl").
    - The Pellet and RacerPRO (via Dig 1.1) reasoning engines were run at both ontology levels using Protégé 3.4, and
    - The Pellet and FaCT++ reasoning engines were run in Protégé 4 for the domain-specific ontology level (i.e., GBR.owl).
- The performance of the inference rules was tested at the application rules ontology (GBR_Rules.owl) with the Jess Inference engine via the SWRL Tab in Protégé 3.4.

6.2.2.3. The Scenario Variables

The scenario variables that can be changed for each test consisted of the data-type and object properties asserted to each instance. The data-type properties available for assertion were: SST (average, maximum and minimum), date, time, LMST, LMSM, MMM, Longitude and Latitude, PAR, rainfall, CO_2, pH, alkalinity, salinity, water depth, turbidity, light quanta, cloud cover, spatial resolution, spatial instrument ID, sensor ID, percent of coral coverage, percent of algal coverage and the fast growth check. The "type of human influence" and "has part" properties are the asserted object properties used in the tests. The "type of human influence" property linked

instances of the "Human Influence" ontology with a reef instance, which in this case was the population information. The "has part" property linked individuals from the taxonomy classes (e.g., coral, algae, plankton, etc.) to a specific coral reef instance to define the species and the community composition it contains. To maintain consistency in the performance tests, eleven "has part" properties were asserted per reef instance with a selection of coral species found in that region. The scenario variables available for manipulation (Table 6.3) were as follows:

- The number of reefs – zero, 3 or 4;
- The data collection intervals (half-hourly versus daily);
- Number of property values asserted to each instance – zero, 13 or 26; and
- The number of atoms in each SWRL inference rule.

6.2.2.4. The Scenario Parameters

The scaling of triples versus performance was the focus of the first group of assessments. As detailed in Chapter 2, a triple is a three component statement to describe an entity (i.e., subject, predicate and object). The assessment scenarios had two means to increase the number of triples; firstly, by the quantity of reef instances created and secondly, by the number of properties asserted to each instance. The number of triples was increased by manipulating the following attributes:

- Additional reef instances – number of reefs (3 or 4);
- Additional reef instances – daily versus half-hourly intervals;
- Additional property assertions – SST only, which required 13 property assertions due to the community composition assertions, versus all values asserted to each individual reef instance (26 assertions).

The time to run the inference rules with a growing number of triples was the focus of the second set of assessments. SWRL rules use propositional logic in the form of Horn-like rules that comprise conjunctions of atoms (refer Chapter 2). The inference rules from Chapters 4 and 5 (i.e., the coral bleach indices) were applied in these tests with the Jess inference engine, to infer conclusions with a growing number of triples. The following attributes were manipulated to compare triple quantity versus inferencing time performance (Table 6.3):

- A growth in triples via additional reef instances – number of reefs (3 or 4);
- A growth in triples via additional properties – 13 property assertions (SST only) versus all values asserted (26 assertions); and

- An increase in the number of atoms that comprise the SWRL rules (5, 9 or 16 atoms).

6.3. Results and Discussion

6.3.1. Limitations

The Java Virtual Machine (JVM) on Windows client operating system (32-bit) is limited to a maximum of 1.6 GB of memory. Because the Protégé framework and the Kepler workflow system are Java-based applications, the scalability of the Semantic Reef system in this client-side desktop environment was constrained. Accordingly, there was a distinct linear progression in the required JVM heap space memory versus the quantity of triples. The following lists the memory allocation events that occurred and the actions taken, which were required to complete the performance analysis tasks:

- The memory allocation for the Java heap space was extended to the maximum which allowed the taskes to be completed.
- RacerPRO, at the 10,000 triple mark, and FaCT++, at the 250,000 triple mark was not able to complete the reasoning process. Therefore, in the final trials, the Pellet reasoning engine in both Protégé 3.4 and Protégé 4 was employed to quantitatively measure the time to reason over the KB.

The RacerPRO reasoning engine which was executed through the DIG interface in Protégé 3.4 and 4 was not able to handle a large number of triples. The expressivity offered by DIG 1.1 is insufficient to capture general OWL-DL ontologies because it does not recognise data-type properties (Bechhofer 2006) and therefore any axioms based on data-type properties to infer knowledge are ignored. For example, the Boolean data-type property “is hermatypic” or the float data-type properties “has longitude/latitude” are important axioms when classifying the KB by bleaching susceptibility or location reef-type. Hence, as RacerPRO is implemented via the DIG 1.1 the inability to continue with its use was not detrimental to the main goals and functionality of the application.

6.3.2. Loading and Reasoning Functionality – Results

The time to load triples to the KB and reason over them was the initial focus. To begin, tests were run five times repeatedly to establish a consistent pattern. Because there was minimal variation between the replicates the performance tests were trialled with three runs each and Table

6.4 depicts the averaged results. The number of total instances and triples in the KB and the resulting time in seconds taken to run the reasoning engines were the key factors. Notably, the lack of data for the RacerPRO reasoner and the incomplete data for the FaCT++ reasoner was due to the memory allocation errors in handling larger numbers of triples.

Coral_Reef_GBR.owl								
REASONER TEST 3 MONTHS	*Legend*	*Instances*	*Triples*	*Load Triples (s)*	*Load KB (s)*	*Protégé 3.4 Pellet (s)*	*Protégé 4 Pellet (s)*	*Protégé 4 FaCT++ (s)*
N0 ASSERTED INSTANCES	**A**	67	160	8.59	10.78	11.53	2.70	83.65
3 REEFS/ DAILY / SST ONLY	**B**	337	5400	16.17	18.83	19.45	172.00	157.97
3 REEFS/ DAILY / ALL VALUES	**C**	337	10000	26.88	29.46	29.73	261.99	527.39
3 REEFS/ HALF-HOURLY/ SST ONLY	**D**	12886	250000	344.53	397.81	734.14	7746.09	
3 REEFS/ HALF HOURLY/ ALL VALS	**E**	12886	440000	772.03	831.17	1347.45	15993.75	
4 REEFS/ HALF HOURLY/ SST ONLY	**F**	17159	330000	467.11	543.13	1003.90	10754.92	
4 REEFS/ HALF HOURLY/ ALL VALS	**G**	17159	590000	1061.02	1157.42	3755.30	27372.18	

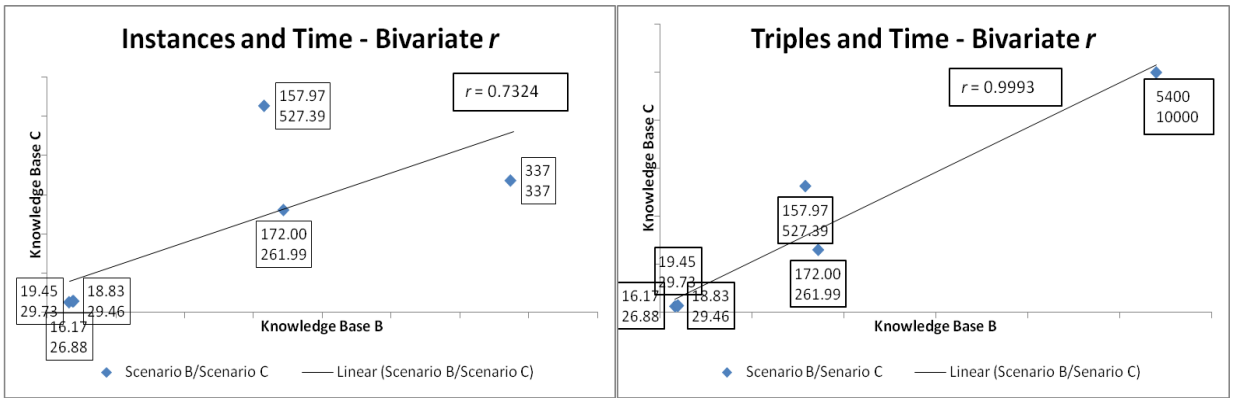
Table 6.4 – KB versions and legend – The test results for quantity of triples versus time to load KB and run the reasoners

Seven versions of the KB were used for the trials (Table 6.4). The first version, labelled A in the legend, which is empty of any reef instances, was included as a "ground-zero" benchmark for the purpose of comparison. Then, the following six tests, labelled B through to G, changed the composition of the KB by the attributes mentioned in the methodology section (§6.2.2.4); specifically, the numbers of reef instances contained in the KB, the number of properties asserted and the temporal intervals of each reef instance.

Statistical results of four scenarios (refer to the scenario parameters §6.2.2.4) which were chosen for a non-biased comparison of the KB versions are shown in Table 6.5. The legend of each KB version in Table 6.4 is indicated in Table 6.5 to depict the comparison operands of the four scenarios. To illustrate, scenario 2 compares the number of data-type property assertions for each reef instance; specifically, the SST values asserted versus all environmental property values. The legend in Table 6.5 refers to the designated test pairs from Table 6.4. Hence, the combinations of KB versions: B and C (Figure 6.1), D and E and F and G were compared in the scenario 2 analysis.

Legend	Marginal Percentage[40] increase/ decrease (%)							Correlation Coefficient Triples vs. Time	Correlation Coefficient Instances vs. Time
	Instances	*Triples*	*Loading Triples (s)*	*Load KB (s)*	*Protégé 3.4 Pellet (s)*	*Protégé 4 Pellet (s)*	*Protégé 4 FaCT++ (s)*		
Scenario 1 - Compare KB (no Reef Instances) to growth in triples via quantity and property assertion									
A&B	80.12	97.04	46.9	42.7	40.7	98.4	47.0	0.886	0.651
A&C	80.12	98.40	68.0	63.4	61.2	99.0	84.1	0.898	0.848
A&D	99.48	99.94	97.5	97.3	98.4	100.0	99.7	0.997	0.758
A&E	99.48	99.96	98.9	98.7	99.1	100.0	99.9	0.997	0.377
A&F	99.61	99.95	98.2	98.0	98.9	100.0	99.9	0.997	0.742
A&G	99.61	99.97	99.2	99.1	99.7	100.0	99.9	0.996	0.220
Scenario 2 – Amount of property assertions – Average SST versus All property values									
*B&C	0.00	46.0	39.8	36.1	34.6	34.3	70.0	*1.00	*0.730
D&E	0.00	43.2	55.4	52.1	45.5	51.6	45.1	1.00	0.890
F&G	0.00	44.1	56.0	53.1	73.3	60.7	44.6	1.00	0.815
Scenario 3 – Amount of reef instances – Temporal intervals - Daily versus Half hourly									
B&D	97.38	97.8	95.3	95.3	97.4	97.8	99.5	1.00	0.526
C&E	97.38	97.7	96.5	96.5	97.8	98.4	99.1	1.00	0.947
Scenario 4 – Amount of triples –SST only and All property values- 3 reefs versus 4 reefs									
D&F	24.90	24.2	26.2	26.8	26.9	28.0	44.1	1.00	1.00
E&G	24.90	25.4	27.2	28.2	64.1	41.6	43.6	1.00	0.984

Table 6.5 – The marginal percentage and correlation coefficients for the four comparison scenarios from the reasoner tests. The results show a correlation between the number of triples versus the time to load and reason over the KB (*an example graph of the Correlation Coefficient for the B&C comparison is depicted in Figure 6.1).



(a) **(b)**

Figure 6.1 – Correlation Coefficient example depicts the comparative relationship of Scenario 2 between KB version B (3 reefs, SST only) and KB version C (3 reefs, all environment values asserted): (a) Instances and time (r =.73), (b) Triples and time (r =1)

[40] Marginal percentage is factored as: ((x-y)/x)*100

The correlation coefficient and the marginal percentage are shown for each scenario. The correlation coefficient analysis looks at bivariate sets of data that compare the test outcome of the KB versions (e.g., A, B, C, etc.) and the change in processing time versus either the number of instances or the number of triples in the KB. Figure 6.1 shows a scatter plot diagram of one correlation comparison: the relationship of KB B and C in scenario 2 (Table 6.5), firstly for instances versus time to process and then triples versus time. This graphic is indicative of the correlation coefficients for all scenarios in Table 6.5. Notably, there was a strong correlation relationship between the duration and the rise in triple quantity and the correlation was weaker for the number of instances asserted to the KB versus processing time (Table 6.5). The increase or decrease in marginal percentage showed a linear progression in scale.

6.3.3. The Loading, Reasoning and Inference Functionality Results

The application ontology ("GBR_Rules.owl") with the SWRL inference rules was assessed in this next series of scenarios. Here, the processing time for loading the KB, reasoning and running inference rules versus the number of triples and reef instances was the significant relationship in the methodology. Firstly, the time involved in the loading and reasoning processes was logged and then the inference rules used in the validation exercise in Chapter 4, which mimic the bleaching indices, were run and the times taken to complete each inference were noted. After the inference rules asserted instances to the relevant "bleach watch" classes, the reasoning trials were then repeated and the times taken for processing were again logged. Table 6.6 depicts the times logged for the pre-rule reasoning, running the inference rules and the post-rule reasoning for the different versions of the KB (i.e., A1 to G1). The distinct linear relationship between time and a growth in triples was observed.

The inference tests were conducted on the Protégé 3.4 framework due to its support for SWRL rules. The rules are defined in the application ontology, which imports all lower ontologies including the domain-specific ontology that contains the instance data. Similar to the previous analyses, Table 6.6 depicts the correlations in the number of triples versus the time to load the KB and reason over the KB. Further, the time to load the SWRL rules, classes, instance data and property assertions to the Jess engine for inference, is also accounted for in Table 6.6. The rules were consistent across all tests and comprised the twelve bleaching indices: SST plus, categories one to five, HotSpot categories one to three, $_{Max}$SST, categories one to three, and one rule with all combined.

GBR_Rules.owl - Pre Rules								
INFERENCE TEST 3 MONTHS	*Legend*	*Instances*	*Triples*	*Load Triples (s)*	*Load KB (s)*	*Protégé 3.4 Pellet (s)*	*Load rules to Jess (s)*	*Inferred Instances*
N0 ASSERTED INSTANCES	**A**	67	250	12.24	17.76	12.03	1.00	0
3 REEFS/ DAILY / SST ONLY	**B**	337	20000	30.21	47.76	20.00	12.87	263
3 REEFS/ DAILY / ALL VALUES	**C**	337	25000	44.22	62.48	29.87	12.87	263
3 REEFS/ HALF HOURLY/ SST ONLY	**D**	12886	265000	434.84	515.68	804.13	555.00	12906
3 REEFS/ HALF HOURLY/ ALL VALS	**E**	12886	455000	997.50	1089.74	1727.50	595.00	12906
4 REEFS/ HALF HOURLY/ SST ONLY	**F**	17159	345000	584.43	693.02	1031.30	749.00	17179
4 REEFS/ HALF HOURLY/ ALL VALS	**G**	17159	605000	1355.89	1486.36	4639.67	749.00	18419
GBR_Rules.owl - Post Rules								
INFERENCE TEST 3 MONTHS	*Legend*	*Instances*	*Triples*	*Load Triples (s)*	*Load KB (s)*	*Protégé 3.4 Pellet (s)*	*Load rules to Jess (s)*	*Inferred Instances*
N0 ASSERTED INSTANCES	**A1**	67	250	12.24	17.76	12.03		
3 REEFS/ DAILY / SST ONLY	**B1**	337	20000	30.67	48.86	21.56		
3 REEFS/ DAILY / ALL VALUES	**C1**	337	25000	44.01	62.66	32.10		
3 REEFS/ HALF HOURLY/ SST ONLY	**D1**	12886	265000	449.53	578.60	816.53		
3 REEFS/ HALF HOURLY/ ALL VALS	**E1**	12886	470000	1028.75	1171.10	2246.10		
4 REEFS/ HALF HOURLY/ SST ONLY	**F1**	17159	360000	603.68	813.75	1119.50		
4 REEFS/ HALF HOURLY/ ALL VALS	**G1**	17159	625000	1375.55	1626.33	3535.15		

Table 6.6 – Inference test legend – The tests results for quantity of triples versus time to load KB and run the reasoner and inference engines

The correlation coefficient and the marginal percentage are shown for each scenario in Table 6.7. The results are similar to the tests from the previous section that were trialled at the domain-specific ontology level (i.e., the "GBR" instance ontology); there was a strong linear correlation between the change in processing time and the quantity of triples.

Legend	***Marginal percentage increase/ decrease***							***Correlation Coefficient Triples vs. Time***	***Correlation Coefficient Instances vs. Time***
	Instances	*Triples*	*Load Triples (s)*	*Load KB (s)*	*Protégé 3.4 Pellet*	*Load rules to Jess*	*Inferred Instances*		
Scenario 1 - Compare Base KB (no Instances) to growth in triples via quantity and property assertion									
Pre Rules									
A&B	80.12	98.75	59.5	62.8	39.8	92.20	100	0.997	0.620
A&C	80.12	99.00	72.32	71.58	59.71	92.23	100	0.997	0.631
A&D	99.48	99.91	97.19	96.56	98.50	99.82	100	0.994	0.472
A&E	99.48	99.95	98.77	98.37	99.30	99.83	100	0.996	0.479
A&F	99.61	99.93	97.91	97.44	98.83	99.87	100	0.994	0.472
A&G	99.61	99.96	99.10	98.81	99.74	99.87	100	0.996	0.436
Post Rules									
A1&B1	80.12	98.75	60.10	63.66	44.20			1.000	0.999
A1&C1	80.12	99.00	72.19	71.66	62.51			1.000	0.999
A1&D1	99.48	99.91	97.28	96.93	98.53			1.000	0.994
A1&E1	99.48	99.95	98.81	98.48	99.46			1.000	0.986
A1&F1	99.61	99.93	97.97	97.82	98.93			1.000	0.994
A1&G1	99.61	99.96	99.11	98.91	99.66			1.000	0.981
Scenario 2 – Amount of property assertions – Average SST versus All property values									
Pre Rules									
B&C	0.00	20.00	31.70	23.56	33.04	0.00	0.00	1.000	0.999
D&E	0.00	41.76	56.41	52.68	53.45	6.72	0.00	1.000	0.999
F&G	0.00	42.98	56.90	53.37	77.77	0.00	6.73	1.000	0.988
Post Rules									
B1&C1	0.00	20.00	30.30	22.02	32.82			1.000	1.000
D1&E1	0.00	43.62	56.30	50.59	63.65			1.000	0.997
F1&G1	0.00	42.40	56.11	49.96	68.33			1.000	0.995
Scenario 3 – Amount of reef instances – intervals Daily versus Half hourly									
Pre Rules									
B&D	97.38	92.45	93.05	90.74	97.51	97.68	97.96	0.999	0.982
C&E	97.38	94.51	95.57	94.27	98.27	97.84	97.96	1.000	0.978
Post Rules									
B1&D1	97.38	92.45	93.18	91.56	97.36			1.000	0.996
C1&E1	97.38	94.68	95.72	94.65	98.57			1.000	0.986
Scenario 4 – Amount of triples –SST only and All property values- 3 reefs versus 4 reefs									
Pre Rules									
D&F	24.90	23.19	25.59	25.59	22.03	25.90	24.87	1.000	1.000
E&G	24.90	24.79	26.43	26.68	62.77	20.56	29.93	1.000	0.993
Post Rules									
D1&F1	24.90	26.39	25.53	28.90	27.06			1.000	1.000
E1&G1	24.90	24.80	25.21	27.99	36.46			1.000	0.999

Table 6.7 – The marginal percentage and correlation coefficients for four comparison scenarios from the Inference tests.

6.3.4. The Inference Rules Atomic Quantity Functionality Results

The inference rules and the Jess inference engine is the focus of the last set of performance analyses. The size of the actual SWRL inference rule versus the processing time required was examined in this experiment. Because each rule is made up of a series of atoms, and each atom relates to an asserted property, to a class member or to a SWRL built-in, they require time and processing power to port to the Jess inference engine via the SWRL Bridge.

Three rules were created for the experiment, Rule 1, Rule 2 and Rule 3, which contain atom quantities of 5, 9 and 16 atoms, respectively (Appendix I). Rule 1 and 2 are focused on the bleaching indices, Rule 1 infers the basic SST+ bleach alert and Rule 2 infers an alert based on three temperature anomaly metrics, shown in Table 6.8 and Table 6.9, respectively. The focal environmental property value is the daily SST, which is used inclusively by KB versions B through to G (Table 6.8, 6.9 and 6.10). Notably, versions B, D and F are KB versions that have only the SST value asserted and are not applicable to Rule 3 (Table 6.10).

Rule 3 contains sixteen atoms that refer to all possible environmental property values available in the current KB (e.g., PAR, pH, salinity, etc.). Because the KB versions B, D and F from Rules 1 and 2 do not contain all available environmental values (26 assertions) but only SST alone, they were not relevant here so versions C, E and G were the focus in the analysis of Rule 3 (Table 6.10). Further, the antecedents of the rules are arbitrary figures chosen for this experiment and not meant to depict an actual hypothesis because, due to the lack of available data, many of the property values for each reef instance are either proxy or fictional. Therefore, the data inserted to the KB to fill the environmental property values such as pH, salinity, turbidity, etc., are proxy values.

INFERENCE RULES TEST GBR_Rules.owl	Legend	Instances	Triples	Property Assertion	Load Rule to Jess	Inferred Instances	Correlation Coefficient Triples vs. Time	Correlation Coefficient Instances vs. Time
Rule 1 - Appendix I								
3 REEFS/ DAILY / SST ONLY	**B**	270	20000	540	1.63	263		
3 REEFS/ DAILY / ALL VALS	**C**	270	25000	540	1.77	263		
3 REEFS/HLF-HRLY/SST ONLY	**D**	12819	265000	25638	39.80	12906		
3 REEFS/HLF-HRLY/ALL VALS	**E**	12819	455000	25638	43.00	12906		
4 REEFS/HLF-HRLY/SST ONLY	**F**	17092	345000	34184	52.93	17179		
4 REEFS/HLF-HRLY/ALL VALS	**G**	17092	605000	34184	54.50	18419		
	B&C	0.00	20.00	0.00	7.55	0.00	1.0000	1.0000
	D&E	0.00	41.76	0.00	7.44	0.00	0.9994	1.0000
Marginal percentage	**F&G**	0.00	42.98	0.00	2.87	6.73	0.9993	0.9990
increase/decrease	**B&D**	97.9	92.5	97.9	95.9	98.0	0.9982	0.9998
	C&E	97.9	94.5	97.9	95.9	98.0	0.9996	0.9997
	D&F	25.0	23.2	25.0	24.8	24.9	1.0000	1.0000
	E&G	25.0	24.8	25.0	21.1	29.9	1.0000	0.9990

Table 6.8 –. The marginal percentage and correlation coefficients for Rule 1 with 5 atoms (refer Appendix I). The number of triples and asserted, or inferred, instances versus the time to load the rules to the Jess inference engine.

INFERENCE RULES TEST GBR_Rules.owl	Legend	Instances	Triples	Property Assertion	Load Rule to Jess	Inferred Instances	Correlation Coefficient Triples vs. Time	Correlation Coefficient Instances vs. Time
Rule 2 - Appendix I								
3 REEFS/ DAILY / SST ONLY	**B**	270	20000	1080	2.20	11		
3 REEFS/ DAILY / ALL VALS	**C**	270	25000	1080	2.30	11		
3 REEFS/HLF-HRLY/SST	**D**	12819	265000	51276	75.57	779		
3 REEFS/HLF-HRLY/ALL VALS	**E**	12819	455000	51276	76.30	779		
4 REEFS/HLF-HRLY/SST	**F**	17092	345000	68368	98.00	779		
4 REEFS/HLF-HRLY/ALLVALS	**G**	17092	605000	68368	98.80	812		
	B&C	0.00	20.00	0.00	4.35	0.00	0.9999	1.0000
	D&E	0.00	41.76	0.00	0.96	0.00	0.9966	1.0000
Marginal percentage	**F&G**	0.00	42.98	0.00	0.81	4.06	0.9962	1.0000
increase/decrease	**B&D**	97.9	92.5	97.9	97.1	98.6	0.9903	1.0000
	C&E	97.9	94.5	97.9	97.0	98.6	0.9977	1.0000
	D&F	25.0	23.2	25.0	22.9	0.0	1.0000	1.0000
	E&G	25.0	24.8	25.0	22.8	4.1	1.0000	1.0000

Table 6.9 - The marginal percentage and correlation coefficients for Rule 2 with 9 atoms (refer Appendix I). The number of triples and asserted, or inferred, instances versus the time to load the rules to the Jess inference engine

INFERENCE RULES TEST GBR_Rules.owl	Legend	Instances	Triples	Property Assertion	Load Rule to Jess	Inferred Instances	Correlation Coefficient Triples vs. Time	Correlation Coefficient Instances vs. Time
Rule 3 - Appendix I								
3 REEFS/ DAILY / ALL VALS	**C**	270	25000	2160	3.67	15		
3 REEFS/HLF-HRLY/ALL VALS	**E**	12819	455000	102552	147.30	2915		
4 REEFS/HLF-HRLY/ALL VALS	**G**	17092	605000	136736	198.00	2915		
Marginal percentage increase/decrease	**C&E**	97.9	94.5	97.9	97.5	99.5	0.9903	0.9998
	E&G	25.0	24.8	25.0	25.6	0.0	1.0000	1.0000
	C&G	98.4	95.9	98.4	98.1	99.5	0.9901	0.9999

Table 6.10 - The marginal percentage and correlation coefficients for Rule 2 with 16 atoms (refer Appendix I). The number of triples and asserted, or inferred, instances versus the time to load the rules to the Jess inference engine

The change in processing time, which involves loading and running the inference engine, versus the number of triples and the number of reef instances was logged and compared (Table 6.8). A linear scale relationship of the time required for the processing task and the growth in triples and reef instances was observed.

The correlation coefficient was distinctly positive for both time versus triples and time versus reef instances. The linear relationship is similarly independent of the number of triples or the number of instances, as opposed to the dominant correlation of only time versus quantity of triples from the reasoning performance analyses in the previous sections. This outcome is explained by the antecedents in the inference rules, which are constant for each test and port only the relevant property, instance or built-in to the Jess Bridge. The outcome would be similar, independent of the number of triples versus the number of reef instances and, therefore, the time to process would also be comparable.

The time involved in running the reasoning engine and the Inference engine were directly relative to the quantity of triples in the Knowledge Base. At 600,000 triples the system was taking longer periods of time to process; however, it was successfully completing the task. The limitations occurred when the quantity of triples exceeded the JVM maximum heap space because, at the desktop computing level, the JVM memory allocation is finite. Hence, the desktop implementation of the Semantic Reef system could not be applied to billions of triples. Although Protégé has been

tested up to two million triples, it will not be feasible as a desktop application for mass data, but is fully suitable for smaller scale practicality (i.e., hypotheses using small to medium sets of data).

The results show the Semantic Reef system to be scalable as a hypothesis tool but restricted only by the platform deployed. The linear scale of time to load the triples and time to run the reasoning and inference engines is in direct correlation to the change in the quantity of triples and, therefore, the more triples populating the KB the longer the processing time. Importantly, however, the functionality is positive for a large quantity of triples (approximately two million) and would be feasible as a desktop hypothesis-driven tool for posing questions of small to medium scale datasets.

Further, there are functions available with semantic technologies that alleviate the need for replication. As discussed in the previous chapter, to represent many reef systems concurrently proxy reef data are employed and can be representative of numerous models of reef-types simultaneously. The models such as "by proximity", "by climate factors" or "by community composition", among others, can be described within the same KB. Then, once the instances have been automatically classified they would belong to more than one class simultaneously through subsumption, which would alleviate many problems of scale. Explicitly, the reef instance is not duplicated but only inferred to belong to the other "reef-type" class by the reasoner.

6.4. Summary

In this chapter a series of functional performance tests were applied to the Semantic Reef system to determine scalability relative to the quantity of data imported to the KB. The system was assessed and proven to handle hypothesis testing for small to medium scale applications at the desktop computing level. Environmental data, logged at half hourly and daily intervals for four reefs for three months (one summer period) was imported to test the system capabilities on a standard desktop computer. The triples generated were dependent on not only the reef instances but on how many properties were asserted to each instance, which ranged from SST alone to scenarios with all available environmental values.

Then, the quantity of instances and/or the property values asserted to the instances were extended or varied in the testing scenarios. The experiments were to find a correlation in processing time versus either the quantity of instances or the quantity of triples. The size of the KB increased with both instances and triples. However, if the numbers of instances are static, the numbers of triples can still increase as associated property values are asserted to describe each instance. Consequently, the more triples within the system, the longer the reasoning and inference engines would take to complete the computation of the inferred hierarchy.

Because the scaling of the run time is proportional to the number of triples, the KB can conceivably handle years of data for a reef or number of reefs on a 32-bit desktop computer. For example, the version C KB (Table 6.8) contained data on 3 reefs with daily logs for 26 environmental variables over a 90 day period (one summer). Version C of the "GBR_Rules.owl" ontology also included the inference rules which raised the amount of triples. Overall, 270 temporal reef instances were allotted to this version of the KB with 26 environmental properties asserted to each. The result was the production of approximately 25,000 triples. Proportionally, the KB can handle data up to 2 million triples on the 32-bit computer platform, which can equate to changes in the scenarios such as:

- The addition of more reefs (up to approximately 250 reefs);
- The extension of the durations (total annual data versus summer data);
- The change in the data logging intervals (e.g., from daily to hourly);
- The addition of previous years for long term analysis, either total annual data or summer periods (i.e., scenario C could be extended to include 80 previous summers); or
- The addition of other asserted properties for environmental parameters.

To simulate the likely computing platform encountered by a researcher posing random questions over available data, the system was trialled on a desktop computing environment. Therefore, a standard desktop computer, a Windows client operating system, was chosen for the performance analysis. The practical limitations of this platform were associated with the size of the JVM memory allocation. The tools implemented as components of the Semantic Reef system are Java-based and the default Java memory space allocation is 200MB, which proved insufficient as more triples were introduced. The Java memory allocation was extended to overcome the problem and, upon increasing the available memory to the JVM, the system completed the set tasks. However, because the memory limits available for the JVM, the Semantic Reef system is limited by available memory resources.

The results from the performance analyses confirmed the time to load, reason and infer over the KB, versus the quantity of triples, scaled proportionally. There was a distinct linear correlation between the quantity of triples in the KB and the time to complete the processing tasks and so the platform variables of the computer limit the scalability of the Semantic Reef system, not the system itself.

To easily scale to greater triples at the desktop deployment level an alternative to the basic 32-bit operating environment is a 64-bit platform. A 64-bit operating system would alleviate the limitations on JVM memory allocations because the constraints on the available memory have a higher default boundary; thus the capacity for a larger number of triples in the KB would be possible. However, the software components employed in the Semantic Reef architecture (i.e., the KB framework, the reasoning and inference engines and the workflow tools) offered greater support for the 32-bit version at the time of development. Hence, the 32-bit platform was adopted for the initial development to the proof of concept stage and the performance tests.

As a predictive tool the Semantic Reef system is inappropriate to run on a desktop machine for very large data sets (i.e., greater than two million triples). To accomplish broader scale prediction, given the quantity of data soon to be available, a computing paradigm to alleviate the limitations of the desktop computer would be the appropriate platform. Alternatives to the desktop platform are available that would facilitate the scaling of triples, such as a Grid computing paradigm or server-side platform. This would alleviate the constraints imposed by the memory limitations of the desktop paradigm.

In conclusion, as a hypothesis tool, to disclose or extract anomalies, phenomena and knowledge in data from disparate sources, the Semantic Reef system is completely feasible. Most propositions could be processed on a desktop computer with samples of data, imported to develop the rules and hypotheses, for *in situ* observation.

Chapter Seven

Conclusion and Discussion

7.1. Overview

Chapter 7 concludes the thesis with a summary of the research findings. After presenting the Semantic Reef architecture and the validation and evaluation results in the preceding chapters a revisit to the initial hypothesis is timely:

To assess the feasibility of using semantic inference in a hypothesis tool to facilitate research on coral reefs by inferring information and/or knowledge from multi-scale, distributed data.

The functional validation and evaluation of the Semantic Reef architecture, by a series of experimental scenarios, confirmed this research hypothesis entirely.

The research aims and objectives and the specific outcomes of the study are described here. In particular, how Semantic Web and scientific workflow technologies can assist with the creation, capture, integration and utilisation of Web-based data pertinent to the coral reef research domain. First, the research objectives are reviewed and followed by a thesis synopsis to show how these objectives were achieved. This is followed by the research contributions, implications, outcomes and constraints and, finally, directions for potential future work and research concludes the thesis.

7.2. Overview of Objectives and Results

The exponential growth in data, appropriately dubbed the data deluge, is a dilemma faced by modern researchers. The large number of data collection instruments deployed, including those that scale-up (e.g., particle accelerator, synchrotrons, etc.) and those that scale-out (e.g., sensor networks), have resulted in an exponential growth of data. Consequently, bottlenecks in the data processing and analysis phases are arising from this increasing volume of raw data and Web available data because many of the analysis procedures require manual intervention. Researchers are finding it progressively more difficult to take advantage of all the data to inform their studies.

The main focus of this research was to explore a method that could help to alleviate the manual data processes and so enable researchers to study coral reefs more fully.

7.2.1. The Research Objectives

The objectives of the study, as listed in Chapter 1, were as follows:

- To investigate the capabilities and synergies of semantic technologies and scientific workflows as methods for data integration.
- To investigate new means in hypothesis modelling and design to enable marine researchers to make efficient use of the data from new collection efforts such as remotely sensed networks. The new means should allow a new research potential to resolve or answer questions such as the effects of climate change on coral reefs.
- To develop an ontology framework that can be reused for any coral reef and is independent of the line of query, the location and/or the data.
- To bridge and combine complex collective knowledge, which is currently held in various data forms within separate research institutions, into one KB for use in hypotheses-driven research.
- To successfully integrate the emerging Semantic Web technologies with scientific workflows into an architecture which allows marine researchers to flexibly pose observational hypotheses based on a richer source of data and information.

7.2.2. Synchronisation to the Objectives

The confirmation of the feasibility of the Semantic Reef conceptual architecture, the methods employed and the strategies taken have been articulated throughout the thesis.

7.2.2.1. The Capabilities and Synergies of the Technologies

The literature review in Chapter 2 on the e-Research paradigm investigated the relevant *enabling technologies* including a study of the strengths and weaknesses of the Semantic Web technologies and scientific workflow tools currently in development. There are activities in research and development fields exploring solutions to the problems arising from the data deluge, such as bottlenecks in data analysis, and many endeavours apply unique mixtures of technology; the Semantic Reef is such a project.

The technological interpretation, analysis and adaptation of these technologies were extended in Chapter 3, with a methodology for the design of the Semantic Reef KB. There, the concepts of a coral reef ecosystem were modelled in a hierarchy of modular reusable and usable

ontologies. The modular design maintains scalability and reusability for future hypotheses independent of reef type, location and community makeup.

This thesis has shown there are synergies in the mixture of the emerging semantic technologies and scientific workflows to handle a variety of data sources and types and, once integrated to the one KB, it can be reasoned over to infer new knowledge or discover phenomena in the data.

7.2.2.2. Flexible Hypothesis Modelling and Design

A more flexible strategy in hypothesis design is enabled by the Semantic Reef system. The researcher is not required to predetermine the exact hypothesis prior to the population of the KB, which is integral to *flexibility in hypothesis modelling*. When posing questions the researcher is only limited by the information in the KB at that time. However, due to the OWA, new concepts can be added to the KB at any time, which will not compromise the knowledge already encapsulated. Further, questions to be asked of the system may not be known prior to data collection. In fact, the questions may emerge or evolve as new data are introduced to the system, sometimes from seemingly disconnected concepts (e.g., adding the local sales information of fertiliser products, schedules from local fishing clubs, etc.). The researcher is able to adapt a hypothesis as more data becomes available or as ideas grow and/or epiphanies emerge.

Flexibility is also afforded by the support for representative data. That is, data from one reef can be indicative of others based on characteristic models rather than only proximity. Automatic reasoning and classification were applied to show how these models may be integrated into the system. To illustrate the degree of flexibility and the ability the system has to automatically link implicit connections, logical axioms were explicitly expressed to describe the reef types by the thermal sensitivity of the community composition and by location. The flexibility in hypothesis design and automated linkages is described and demonstrated in Chapter 5.

7.2.2.3. A Reusable Ontology Framework for Coral Reef Research

The strategy in ontology engineering, detailed in Chapter 3, aimed at *reuse and flexibility*. The Semantic Reef model was built from the semantic building blocks of a domain expert's functional representation of the concepts of any coral reef. The result was a hierarchy of ontological complexity (i.e., informal to formal ontologies), starting from the "reusable" base level taxonomies to the highest layer "usable" application ontologies.

The KB was reused for each example in the validation described in Chapter 4 and capability demonstrations in Chapter 5. The validation in Chapter 4 used the historic environmental data of four reefs with daily temporal SST over eight summer periods, from 1996 to 2003. Then, for the demonstrations shown in Chapter 5, the KB was repopulated with data from a different group of reefs on the GBR. Notably, although the demonstrations centred on the central GBR region, the KB could be repopulated for another reef system, such as Moorea Island or the Bahamas, simply by creating new "usable" domain-specific and application level ontologies that import the reusable lower ontology hierarchy.

7.2.2.4. Data Integration

Data integration is a major aim of Semantic Web technologies. These technologies offer standards to describe concepts in terms that are understandable by a computer. The addition of computer-readable context to disparate data offers the computer enough information to derive meaning of the data, whether it is from data collection or production instruments or sourced from the Web. Ontologies contain the explicit contextual information, the well-defined descriptions and the relationships of concepts and are the basis of semantic technologies. The ability to define concepts to a computer is seminal to new data integration methods such as described in Chapter 2. Ontologies were created to describe the concept of a coral reef (Chapter 3) at both a generic reusable level and also a domain-specific usable level for hypothesis research across any coral reef regardless of its type or location. The ontology languages offer constructs to state explicit relationships such as equivalencies that assist in the bridging of disparate data sets (e.g., staghorn coral from one dataset is *Acropora* in others). Once the computer can make sense of the data, the data is automatically processable by the computer (demonstrated in Chapters 4 and 5).

Then, independent of data origination, the computer can infer decisions or automate classification to connect hidden links in the data. For example in Chapter 5, once axioms were declared to define a "reef-type" by its location or community composition, all reefs that matched the definitions were automatically subsumed to belong to multiple parent classes by the Pellet reasoning engine.

7.2.2.5. Demonstrate the New Semantic Reef System and the Beneficial Differences to Hypothesis-based Research

The Semantic Reef architecture is a proof of concept that represents an exemplar of future methods for managing rich data sources in more productive ways. The employment of semantic inference within the architecture offers benefits in flexible hypothesis design and differences in

knowledge discovery, which include modularity, reusability and data integration and were illustrated in Chapters 3, 4, 5 and 6.

The substantiation of the Semantic Reef KB was depicted in Chapter 4. The validation entailed a reverse-hypothesis methodology that compared outcome of the KB inference rules with actual historic records and research. The result from this demonstration confirmed the accuracy of the Semantic Reef system for use in hypothesis-driven research by its ability to disclose temporal instances that match premises in a proposition for *in situ* observation. Also, the KB is capable of alerting to coral bleaching events via the unorthodox use of logical inference that mimic standard bleaching indices. Notably, although the use of semantic inference is not the optimal way to calculate coral bleaching alert indices, the exercise was representative of an anticipated use of a fully implemented Semantic Reef system.

The modular ontology design within KB aimed to maximise both reuse and usability simultaneously for a new approach to flexible hypothesis design and was explained in Chapter 3 and demonstrated in Chapters 4 and 5. The methodology in ontology design applied a new hybrid of current methods to achieve a separation of data instances from the concept descriptions. This strategy allows the KB to be populated with data from any coral reef in the world for the purpose of new hypotheses. Then, by simply repopulating the KB with data and information relevant to a new study it can be easily reused for a different line of enquiry in any other coral reef in the world. Also, the number of triples stored in the KB is reduced because only instances required for a hypothesis are imported to the system and thus possible to run effectively at the desktop computing level.

The differences semantic inference has from other KR paradigms for hypothesis-driven research were demonstrated in a series of use-cases in Chapter 5. Differences such as the unstructured nature of semantic logic systems that allow for simpler data integration and the automation of knowledge discovery through inference were portrayed. The semantic technologies were applied to automate much of the analysis and hypotheses processes through inference. Semantic inference allows the latent links in disparate data to be automatically connected by applying reasoning methods to classify the data within the KB. The autonomous extrapolation of phenomena was then possible with DL and propositional logic systems (Chapter 5). These capabilities will help to alleviate the bottlenecks being formed as the data deluge grows.

Finally, a quantitative analysis of the functionality and practicality of the Semantic Reef system was presented in Chapter 6 and proved the practical viability of the system as a hypothesis tool.

7.3. The Outcomes and Contributions

"e-Research refers to the development of, and the support for, information and computing technologies to facilitate all phases of research processes." (Allan et al. 2004)

The Semantic Reef use-case is a subset of broader eco-informatics applications which can help process the imminent deluge of data and reduce data to knowledge. The architecture offers an alternative approach to the development, application and execution of observational hypotheses in the marine domain and has been tested and proven to handle limited quantities of disparate data for a range of propositional suppositions and can extract or disclose phenomena within the data.

The following research outcomes and contributions of the study have been realised:

- The Semantic Reef architecture is a feasible design for a hypothesis tool to enable coral reef researchers.
- The Semantic Reef architecture applies Semantic technologies and scientific workflows to integrate data for the purpose of posing observational hypotheses or inferring alerts in a coral reef domain.
- A set of modular reusable ontologies have been developed that describe a generic coral reef ecosystem. The methodology and development was discussed in Chapter 3 and a paper on this was presented at the Knowledge Representation Ontology Workshop (Myers et al. 2008).
- The accuracy of the Semantic Reef system was demonstrated through a reverse-hypothesis method, which validated the system as a proof of concept. The validation was presented in Chapter 4 and a paper on this was presented at the Environmental Research Event conference (Myers et al. 2007)
- Through a select number of use-case examples, which are the first known in the coral reef domain, the benefits and differences of employing semantic inference in the design was explored and demonstrated (Chapter 5).
- Performance and evaluation tests were conducted and the results documented for the Semantic Reef model and architecture on real, world-like, scenarios in the coral reef domain (Chapter 6).
- The simulation experiments showed the feasibility of the Semantic Reef system as a viable tool for research at the desktop computing level (Chapter 6).

- The architectural design and the engineering methodology are not limited to the coral reef domain. They are suitable for other applications that initiate hypothesis-driven research via *in situ* observation, such as terrestrial ecology.
- The Semantic Reef project is a new application in eco-informatics, which is the combination of multiple environmental datasets and modelling tools used to test ecological hypotheses and derive information. Here the combination of expertise in Ecological disciplines was matched with expertise in the computer sciences to form synergies for data-management and knowledge discovery. The result was the development of a tool that has the potential to solve emerging problems such as the integration of disparate data for synthesis and analysis. The development was established in Chapter 3 and 5 and a paper on this was accepted in the Journal of Applied Artificial Intelligence (Myers et al. 2009)
- This thesis is a contribution to the current literature. It details an example unification of coral reef science and ICT disciplinary studies to offer a modern solution to data management. The exploration to possible solutions to some imminent problems of modern research, such as the bottlenecks in data analysis and synthesis caused by the increasing data deluge, was the focus of the thesis. A paper on this topic was presented at the International Coral Reef Symposium (Myers and Atkinson 2009).

7.4. Constraints and Assumptions

7.4.1. The Lack of Data in a Data Deluge

Ultimately the Semantic Reef Knowledge Base will be able to be filled with relevant data as it becomes available. The data, which would come from a diverse range of Internet-mediated sources, would ideally include such elements as salinity, CO_2, pH levels, nitrogen, anthropogenic influences, bathymetry information and species standing stocks. Knowledge may then be derived from the data by questioning for semantic correlation and analysis through the logic systems.

However, data cannot be employed in a hypothesis if it does not exist and, currently, there are gaps in available data. The data may not exist because it has not been or is not being collected, or the data may be from a closed source and the permissions to use the data are dependent on proprietary policies (i.e., access may not be granted due to the governing directives). If inference over the KB is to be of value the data pertinent to the line of enquiry must be available for import to the KB.

One objective of the Semantic Reef project is to draw out gaps in the current marine data. Even in the midst of a data deluge there are lacunae in the data collected, and decisions of where, when and what types of data to gather are imperative to researchers and data collection managers. The managers of funding resources would benefit from information on data requirements, especially in decisions about the most efficient and cost effective deployment strategy for the data acquisition. The researcher would expose gaps in required data for the line of enquiry as they engage a flexible hypothesis tool, such as the Semantic Reef model, because the arbitrary questions they may wish to pose in order to disclose phenomena may not be possible due to unavailable data.

7.5. Future Work

7.5.1. The Deployment Computing Paradigm – Desktop to Grid

The Semantic Reef system is completely feasible as a hypothesis tool to extract anomalies, phenomena and knowledge in data from disparate sources. At the desktop computing level reasoning over large data sets is not necessarily a requirement; instead, at this level, the researcher may engage the system to create, change and manipulate the questions posed, while determining the hypothesis itself. The use of small to medium data sources to populate the KB is appropriate because the hypothesis may evolve as different data becomes available, or the line of enquiry is modified due to an epiphany.

Conversely, however, deployment in a desktop computing paradigm restricts the scalability of the Semantic Reef architecture. The system scales in a linear progression of quantity of triples versus time to process, as the results of the performance analyses in Chapter 6 show. Hence, the system is limited by the practical computing environment factors, such as the maximum memory allocations and storage quota allowed by the JVM and operating system. Due to these limitations, the system would not be capable of managing the large scale data that are being produced by the remotely sensed observation systems and, consequently, would be restricted as a prediction tool, for example, to forecast coral bleaching events.

Notably, the Semantic Reef concept is not constrained to the current tools employed in the development of the proof of concept. The purpose here was not to test specific semantic software and scientific workflow tools but rather to explore the benefits semantic inference and workflows offer to hypothesis-based research in marine research. The scalability of the Semantic Reef system could be extended in other implementations by employing server based versions of the current

software. Further, to accomplish broader scale prediction a Grid computing paradigm would be appropriate.

The natural environment for electronic scientific workflows is a Grid computing paradigm. The application of the Semantic Reef architecture on a Grid infrastructure would eliminate the limitations placed at the desktop computing level (e.g., memory allocations) due to the availability of the processing power required when scaling to the extensively large data sets. A Grid computing infrastructure is to be implemented as a component in future work. The Grid brings a solid, safe, secure infrastructure that facilitates services to control the sharing of hardware, software and data resources. The Semantic Reef architecture requires availability and/or access to data repositories, computer power for data processing and digital storage to realise its full potential. A Grid Computing paradigm will make possible the added processing power and storage required for scaling to a massive amount of data, both live and static. The result will be more efficient investigation methods of the disparate data streams and data sources for research on coral reef ecosystems.

7.5.2. Quality Assurance of the Data

The Semantic Reef KB has no internal quality checks, but instead assumes the incoming data is already quality assured. Known gaps or possible corruption in the data would be an assumption and a recognised part of a researcher's methodology. The goal of this thesis was to demonstrate the feasibility and potential of a new prototype system via a series of demonstrations, examples and evaluations. The data used for the validation component of this study were the datasets studied in previous research on the 1998 and 2002 coral bleaching events (Chapter 4) and was assumed to be free of corruption (Berkelmans et al. 2004; Maynard 2004). The data ported to the KB to demonstrate the system's functionality in Chapter 5 and the performance analyses in Chapter 6 were not required to be quality checked because the purpose of this data was for use to prove the concept of the system, not a real empirical study.

The addition of quality assurance functionality is part of future work. The Kepler workflow software offers some quality assurance functionality, such as checking for gaps in data sequences, adding a scale of belief as provenance annotations, etc. If the Semantic Reef system is to achieve a complete solution to automate certain data processing tasks and alleviate manual intervention it will need to incorporate a quality assurance mechanism for the data.

7.5.3. Usability

Currently the system is not a portable software solution; that is, is not an implementation. To date, the Semantic Reef architecture is a proof of concept that has utilized emerging technologies as they are being developed. Notably, the concept can be implemented with other available software tools because it is not constrained to the specific tools used for its development. The aim of this study was to assess the viability of semantic inference to assist scientists in automating data analysis during the hypothesis process. This aim has been reached and the concept of a semantically-driven hypothesis tool is a feasible solution to current research problems, such as data integration.

A portal for scientists to ask questions as they arise is a component of future work. The integration of a Human Computer Interface (HCI) would add practical functionality to the Semantic Reef architecture where applications and hypotheses could be tested. A deployment platform is an eminently feasible future feature of the system.

A number of systems, already in development, offer an avenue for the exploration for a web portal interface to the Semantic Reef system. The myExperiment Virtual Research Environment (VRE) supports the sharing of research objects used by scientists, such as scientific workflows (De Roure et al. 2008) and is both a social infrastructure for enabling collaborations and a platform for conducting research. The workflow system in the myExperiment VRE is the Taverna workflow tool, which would require significant changes to incorporate it into the Semantic Reef architecture.

Alternatively, the Hydrant project[41] is a web portal application for the Kepler workflow system that provides the means for users to deploy and share their workflows and results via the Web. Hydrant supports workflow execution and allows users to assign parameters to workflows and permissions to other users. Notably, the Hydrant web portal is an appropriate system for end-user application; however, it is currently still in development as a proof of concept.

Well-managed communications, education and direction in formatting a hypothesis as Horn-like rules, while maintaining monotonicity, would be integral for a turn-key software solution. The Semantic Reef project is an application in eco-informatics, which combines expertise in both the ecological and information technology domains. Open communication was required to bridge disciplinary knowledge gaps, particularly when translating hypotheses from the specifications of domain experts to sets of semantic inference rules. Therefore, collaboration will

[41] www.hpc.jcu.edu.au/hydrant

remain an integral necessity because crossing disciplines to create hypotheses, of the kind described here, is a complex undertaking.

7.5.4. Causal Logics

The Semantic Reef system remains at this stage predominantly forward chaining. Specifically, the system implements only deductive and inductive reasoning methods, which models predictions and evaluations prescribed from the explicit descriptions by domain experts of the coral reef concept. To implement a backward chaining paradigm, such as cause and effect with abductive reasoning, probabilistic logic systems such as Bayesian beliefs, causal logics and modal logics will be required.

Causal research is an important facet of ecological modelling and a requirement for marine research. There is currently development in semantic technologies that are incorporating probability reasoning and are building tools that adopt these abductive methods. An example of these developments is Probabilistic OWL (PROWL), a version of OWL that supports probabilistic logic (Costa and Laskey 2006). However, the endeavours such as PROWL are not in mature stages of development. Future work on the Semantic Reef Project would see the integration of the backward chaining inference for a complete hypothesis system (Costa et al. 2005).

7.6. Final Remarks

To respond to ever changing trends in research there is a resulting need to continually explore new methods and/or adapt current technologies in new and interesting ways to support the research process. As technologies evolve so, too, do the challenges that arise from employing new methods. The new data collection methods that scale-up or scale-out are resulting in a data deluge and the challenge is to find efficient and effective methods to process the available data. The Semantic Reef architecture is one means to overcome the data bottlenecks, by merging technologies to explore their possible synergies to observe how they may help solve current research problems.

The demand for automatic data analysis and hypothesis testing is increasing. This demand will be even more imperative as current and future data production and collection infrastructures such as emerging developments in sensor networks to remotely monitor reef systems are deployed. The Semantic Reef is a new approach to data analysis and interpretation that will be extendable to other areas and issues apart from environmental monitoring and climate change.

While coral reefs are the focus environment of this study, the concepts described in this thesis are applicable to many other disciplines. The ontology hierarchy is flexible and modular and could be adapted or rewritten to describe any domain. The ontologies, which are central components in Knowledge Representation, explicitly define expressions to add computer-readable context to give meaning to the data and can be created to describe any concept. The coral reef domain was the focus to prove the concept of the Semantic Reef project; however, the architecture of the project is interchangeable to any discipline that applies observation hypothesis research.

This thesis was undertaken within a cross-discipline, collaborative environment. Diverse partners and expertise have been combined (including the Australian Institute of Marine Science and the Great Barrier Reef Marine Park Authority). The uptake of new technology relies on the involvement of end users and new ICT developments will only succeed if new technologies can be applied to solving end user problems. This dissertation has shown that semantic technologies can successfully be applied to hypothesis-driven research in marine science.

Bibliography

Access-Economics 2005, 'Measuring the economic and financial value of the Great Barrier Reef.', Great Barrier Reef Marine Park Authority, Townsville, Australia.

Adamic, L. & Huberman, B. 2002, 'Zipf's law and the Internet', *Glottometrics*, vol. 3, no. 1**,** pp. 143-150.

AIMS 2007a, *Chlorophyll monitoring on the Great Barrier Reef*, Australian Institute of Marine Science, viewed 15th April 2009, http://www.aims.gov.au/docs/data-centre/chlorophyllmonitoring.html.

AIMS 2007b, *Threats to coral reefs*, Australian Institute of Marine Science, viewed 29th January 2009, http://www.aims.gov.au/docs/research/biodiversity-ecology/threats/threats.html.

AIMS 2008, *Marine blueprint - climate change and the fate of the Great Barrier Reef*, Australian Institute of Marine Science, viewed 14th April 2009, http://www.aims.gov.au/docs/research/climate-change/position-paper.html.

AIMS 2009, *The Long Term Monitoring Program (LTMP)*, Australian Institute of Marine Science, viewed 30th June 2009, http://www.aims.gov.au/docs/research/monitoring/monitoring.html.

Alabri, A., Hunter, J., Ingen, C. V. & Abal, E. 2009, 'The Health-e-Waterways project: data integration for smarter collaborative whole-of-water cycle management', *International Conference on Complex, Intelligent and Software Intensive Systems (CISIS 2009)*, Fukuoka, Japan, March 16-19.

Allan, R., Allden, A., Boyd, D., Crouchley, R., Harris, N., Lyon, L., Robiette, A., De Roure, D. & Wilson, S. 2004, 'Roadmap for a UK virtual research environment', *Report of the JCSR VRE Working Group*, Joint Information Systems Committee (JISC), London, England.

Allemang, D. & Hendler, J. 2008, Semantic Web for the working ontologist: effective modeling in RDFS and OWL, Morgan Kaufmann, Burlington, MA, USA.

Altintas, I., Berkley, C., Jaeger, E., Jones, M., Ludäscher, B. & Mock, S. 2004, 'Kepler: an extensible system for design and execution of scientific workflows', *Proceedings of the 16th International Conference on Scientific and Statistical Database Management (SSDBM 04)*, Santorini Island, Greece, June, IEEE, pp. 423- 424.

Antoniou, G., Billington, D., Governatori, G. & Maher, M. J. 2001, 'Representation results for defeasible logic', *ACM Trans. Comput. Logic*, vol. 2, no. 2**,** pp. 255-287.

Antoniou, G. & van Harmelen, F. 2004, A Semantic Web primer The MIT Press, Cambridge, MA, USA.

Antoniou, G. & van Harmelen, F. 2008, A Semantic Web primer (2nd edition), The MIT Press, Cambridge, MA, USA.

Atkinson, I., Buckle, A. M., Groenewegen, D., Nicholas, N., Treloar, A. & Beitz, A. 2008, 'ARCHER – an enabler of research data management', *4th IEEE International Conference on eScience (e-Science 08)*, Indianapolis, IN, USA, 7-12 December, IEEE International, pp. 246-252.

Baader, F., Calvanese, D., McGuinness, D. L., Nardi, D. & Patel-Schneider, P. F. (Eds.) 2003, *The Description Logic handbook: theory, implementation and applications*, Cambridge University Press, Cambridge, USA.

Baader, F., Calvanese, D., McGuinness, D. L., Nardi, D. & Patel-Schneider, P. F. (Eds.) 2007, *The Description Logic handbook theory, implementation and applications (2nd ed.)*, Cambridge University Press, Cambridge, USA.

Baader, F., Horrocks, I. & Sattler, U. 2005, 'Description logics as ontology languages for the Semantic Web ', in Hutter, D. & Stephan, W. (Eds.), *Mechanizing mathematical reasoning*, Springer-Verlag, Berlin/Heidelberg, Germany, pp. 228-248.

Bechhofer, S. 2006, *DIG 2.0: The DIG Description Logic interface*, University of Manchester DIG Working Group, viewed 17th July 2009, http://dig.cs.manchester.ac.uk/index.html.

Bechhofer, S., Möller, R. & Crowther, P. 2003, 'The DIG Description Logic interface', *Proceedings of the International Workshop on Description Logics (DL2003)*, Rome, Italy, 5-7 September, CEUR.

Bergman, M. 2001, 'The Deep Web: surfacing hidden value', *Journal of Electronic Publishing*, vol. 7, no. 1**,** pp. 1-20.

Berkelmans, R. 2002, 'Time-integrated thermal bleaching thresholds of reefs and their variation on the Great Barrier Reef', *Marine Ecology Progress Series*, vol. 229, no.**,** pp. 73-82.

Berkelmans, R. 2009, 'Bleaching and mortality thresholds: how much is too much?', in van Oppen, M. J. H. & Lough, J. M. (Eds.), *Coral bleaching - patterns, processes, causes and consequences*, Springer, Berlin/Heidelberg, Germany, pp. 103-19.

Berkelmans, R., De'ath, G., Kininmonth, S. & Skirving, W. J. 2004, 'A comparison of the 1998 and 2002 coral bleaching events on the Great Barrier Reef: spatial correlation, patterns, and predictions', *Coral Reefs*, vol. 23, no. 1**,** pp. 74-83.

Berkelmans, R., Done, T., Goggin, L. & Harriott, V. 2002, *Coral bleaching and global climate change - current state-of-knowledge*, CRC Reef Research Centre Ltd, viewed 29th March 2009, http://www.reef.crc.org.au/publications/brochures/bleaching_brochure.pdf.

Berners-Lee, T. 1990, 'WorldWideWeb: proposal for a HyperText project', Hypertext project proposal, presented at CERN, 12 November, 1990, Available online at http://www.w3.org/Proposal.

Berners-Lee, T. 1999, *Web Architecture from 50,000 feet*, W3C, viewed 16th May. 2004, http://www.w3.org/DesignIssues/Architecture.html.

Berners-Lee, T. 2000a, 'Semantic Web on XML', Keynote Speech, presented at the XML 2000, Washington, DC, USA, 6 December, Available online at http://www.w3.org/2000/Talks/1206-xml2k-tbl.

Berners-Lee, T. 2000b, Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web, Collins.

Berners-Lee, T. 2002, 'The World Wide Web - past present and future: exploring universality', The Commemorative Lecture 2002, Japan Prize, presented at Narita, Japan, Available online at http://www.w3.org/2002/04/Japan/Lecture.html.

Berners-Lee, T. 2003, 'WWW past, present and future - lessons from building the Web applied to building the Semantic Web', Keynote Speech, presented at the Royal Society London, England, 22 September, Available online at http://www.w3.org/2003/Talks/0922-rsoc-tbl/.

Berners-Lee, T. 2007, *Linked data*, W3C, viewed 29th August 2009, www.w3.org/DesignIssues/LinkedData.html.

Berners-Lee, T. 2008, *The scale-free nature of the Web*, W3C, viewed 15th March 2009, http://www.w3.org/DesignIssues/Fractal.html.

Berners-Lee, T., Fielding, R. & Masinter, L. 2005, *Uniform Resource Identifier (URI): generic syntax*, Network Working Group, viewed 29th January 2009, http://www.gbiv.com/protocols/uri/rfc/rfc3986.html.

Berners-Lee, T., Hall, W., Hendler, J., Shadbolt, N. & Weitzner, D. J. 2006, 'Creating a Science of the Web', *Science*, vol. 313, no. 5788**,** pp. 769-771.

Berners-Lee, T., Hall, W., Hendler, J. A., O'Hara, K., Shadbolt, N. & Weitzner, D. J. 2006, 'A framework for Web Science', *Found. Trends Web Sci.*, vol. 1, no. 1**,** pp. 1-130.

Berners-Lee, T., Hendler, J. & Lassila, O. 2001, 'The Semantic Web', *Scientific American*, vol. 284, no. 5**,** pp. 34-43.

Boley, H. & Kifer, M. 2008, *RIF basic logic dialect*, W3C Working Draft, viewed 17th February 2009, http://www.w3.org/TR/rif-bld/.

BOM 2008, *Australian Bureau of Meteorology - tropical cyclone outlooks*, Australian Bureau of Meteorology, viewed 20th April 2009, http://www.bom.gov.au/weather/cyclone/tc-outlooks.shtml.

Brachman, R. J. & Levesque, H. J. 2004, Knowledge representation and reasoning, Morgan Kaufmann, San Francisco, CA, USA.

Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E. & Yergeau, F. 2008, *Extensible Markup Language (XML) 1.0 (5th edition)*, W3C Recommendation, viewed 20th February 2009, http://www.w3.org/TR/xml/.

Brickley, D., Guha, R. V. & McBride, B. 2004, *RDF vocabulary description language 1.0: RDF Schema*, W3C Recommendation, viewed 21st February 2009, http://www.w3.org/TR/rdf-schema/.

Brown, B. E. 1997, 'Coral bleaching: causes and consequences', *Coral Reefs*, vol. 16, , no. Supplement 1**,** pp. S129-S138.

Catlett, C. 2002, 'The philosophy of TeraGrid: building an open, extensible, distributed TeraScale facility', *Proceedings of the 2nd IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGRID 02)*, Berlin, Germany, 20-24 May, IEEE Computer Society, pp. 8-9.

CC 2009, *Creative Commons*, viewed 15th August 2009, http://creativecommons.org/.

CERN 2008, *Large Hadron Collider*, European Organization for Nuclear Research (CERN), viewed 15th March 2009, https://lhc2008.web.cern.ch/LHC2008/.

Chandrasekaran, B., Josephson, J. & Benjamins, V. 1999, 'What are ontologies, and why do we need them?', *Intelligent Systems and Their Applications, IEEE*, vol. 14, no. 1**,** pp. 20-26.

Clark, B. R., Godfray, H. C. J., Kitching, I. J., Mayo, S. J. & Scoble, M. J. 2009, 'Taxonomy as an e-Science', *Phil. Trans. R. Soc. A*, vol. 367, no. 1890**,** pp. 953-966.

Corcho, O., Alper, P., Kotsiopoulos, I., Missier, P., Bechhofer, S. & Goble, C. 2006, 'An overview of S-OGSA: a reference Semantic Grid architecture', *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 4, no. 2**,** pp. 102-115.

Corcho, O., Fernández-López, M. & Gómez-Pérez, A. 2003, 'Methodologies, tools and languages for building ontologies: where is their meeting point?', *Data Knowl. Eng.*, vol. 46, no. 1**,** pp. 41-64.

Costa, P. C. G. & Laskey, K. B. 2006, 'PR-OWL: a framework for probabilistic ontologies', *4th International Conference on Formal Ontology in Information Systems (FOIS 06)*, Baltimore, MD, USA, 9-11 November.

Costa, P. C. G., Laskey, K. B., Laskey, K. J. & Pool, M. 2005, 'PR-OWL: a bayesian ontology language for the Semantic Web ', *Proceedings of the ISWC Workshop on Uncertainty Reasoning for the Semantic Web (URSW 05)*, CEUR-WS, Galway, Ireland, vol 173, pp. 88-107.

CRC 2006, *Coral Reefs*, CRC Reef Research Centre, viewed 16th March 2006, http://www.reef.crc.org.au/discover/coralreefs/index.html.

CREON 2008, *Coral Reef Environmental Observatory Network*, viewed 20th October 2009, http://www.coralreefeon.org/.

CSIRO 2007, *Reef Temp*, GBRMPA, viewed 20th October 2009, http://www.cmar.csiro.au/remotesensing/reeftemp/web/ReefTemp.htm.

Czajkowski, K., Fitzgerald, S., Foster, I. & Kesselman, C. 2001, 'Grid information services for distributed resource sharing', *Proceedings of the 10th IEEE Symposium on High Performance Distributed Computing*, IEEE, vol 7513, pp. 181-194.

Davey, M., Holmes, G. & Johnstone, R. 2008, 'High rates of nitrogen fixation (acetylene reduction) on coral skeletons following bleaching mortality', *Coral Reefs*, vol. 27, no. 1**,** pp. 227-236.

DBpedia 2009, *The DBpedia knowledge base*, viewed 30th September 2009, http://wiki.dbpedia.org/.

de Kunder, M. 2006, 'Daily estimated size of the World Wide Web', MSc, Tilburg University, Warandelaan, The Netherlands.

De Roure, D., Gil, Y. & Hendler, J. 2004, 'Guest editors' introduction: e-Science', *IEEE Intelligent Systems*, vol. 19, no. 1**,** pp. 24-25.

De Roure, D. & Goble, C. 2009, 'Software design for empowering scientists', *IEEE Software*, vol. 26, no. 1**,** pp. 88-95.

De Roure, D., Goble, C., Bhagat, J., Cruickshank, D., Goderis, A., Michaelides, D. & Newman, D. 2008, 'myExperiment: defining the social virtual research environment', *4th IEEE International Conference on eScience (e-Science 08)*, Indianapolis, IN, USA, 7-12 December, IEEE Press.

De Roure, D., Goble, C. & Stevens, R. 2007, 'Designing the myExperiment virtual research environment for the social sharing of workflows', *Proceedings of the Third IEEE International Conference on e-Science and Grid Computing (e-Science 2007)* Bangalore, India, 10-13 December, IEEE Computer Society, pp. 603-610.

De Roure, D., Goble, C. & Stevens, R. 2009, 'The design and realisation of the myExperiment virtual research environment for social sharing of workflows', *Future Generation Computer Systems*, vol. 25, no.**,** pp. 561-567.

De Roure, D. & Hendler, J. A. 2004, 'E-Science: the grid and the Semantic Web', *Intelligent Systems, IEEE*, vol. 19, no. 1**,** pp. 65-71.

De Roure, D., Jennings, N. R. & Shadbolt, N. R. 2003, 'The Semantic Grid: a future e-Science infrastructure', in Berman, F., Fox, G. & Hey, A. J. G. (Eds.), *Grid computing - making the global infrastructure a reality*, John Wiley and Sons Ltd., West Sussex, England, pp. 437-470.

De Roure, D., Jennings, N. R. & Shadbolt, N. R. 2005, 'The Semantic Grid: past, present, and future', *Proceedings of the IEEE*, vol. 93, no. 3**,** pp. 669-681.

Done, T., Harriott, V., Berkelmans, R. & Goggin, L. 2005, *Coral bleaching and global climate change - current state-of-knowledge (2nd edition)*, CRC Reef Research Centre Ltd, viewed 17th March 2009, http://www.reef.crc.org.au/publications/brochures/Bleachingfront.htm.

EGEE 2009, *Enabling Grids for e-Science*, viewed 20th February 2009, http://project.eu-egee.org/.

FaCT++ 2008, *Fast Classification of Terminologies Description Logic classifier*, School of Computer Science, University of Manchester, viewed 29th September 2008, http://owl.man.ac.uk/factplusplus/.

Fensel, D., Hendler, J. A., Lieberman, H. & Wahlster, W. 2002, Spinning the Semantic Web: bringing the World Wide Web to its full potential, The MIT Press, Cambridge, MA, USA.

Fernandez-Lopez, M. & Gomez-Perez, A. 2002, 'Overview and analysis of methodologies for building ontologies', *The Knowledge Engineering Review*, vol. 17, no. 02**,** pp. 129-156.

Foster, I. 2002, 'The Grid: a new infrastructure for 21st century science', *Physics Today*, vol. 55, no. 2**,** pp. 42-47.

Foster, I., Jennings, N. R. & Kesselman, C. 2004, 'Brain meets brawn: why Grid and agents need each other', *Proceedings from the 3rd International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 04)*, IEEE, New York, NY, USA, vol 01, pp. 8-15.

Foster, I. & Kesselman, C. 1998, 'The Grid: Blueprint for a Future Computing Infrastructure.', *Chapter 2*,, Morgan Kaufmann Publishers, pp.

Foster, I., Kesselman, C., Nick, J. M. & Tuecke, S. 2002a, 'Grid services for distributed systems integration', *IEEE Computer*, vol. 35, no. 6**,** pp. 37-46.

Foster, I., Kesselman, C., Nick, J. M. & Tuecke, S. 2002b, 'The physiology of the Grid: an open grid services architecture for distributed systems integration', *Open Grid Service Infrastructure WG*, Global Grid Forum (GGF), Chicago, IL, USA.

Foster, I., Kesselman, C. & Tuecke, S. 2001, 'The anatomy of the Grid: enabling scalable virtual organizations', *Int J Supercomputer Appl*, vol. 15, no. 3**,** pp. 200-222.

Frey, J. G., De Roure, D. & Carr, L. A. 2002, 'Publication at source: scientific communication from a publication web to a data grid', *Euroweb 2002 Conference, The Web and the GRID: from e-science to e-business*, Oxford, UK, 17-18 December, BCS.

Frey, J. G., Hughes, G. V., Mills, H. R., schraefel, m. c., Smith, G. M. & De Roure, D. 2004, 'Less is more: lightweight ontologies and user interfaces for smart labs', *Proceedings of the UK e- Science All Hands Meeting*, EPSRC, Nottingham, UK, vol 1, pp. 8-16.

Friedman-Hill, E. 2003, Jess in action : Java rule-based systems (in action series), Manning Publications, Greenwich, CT, USA.

Garcia, A., O'Neill, K., Garcia, L. J., Lord, P., Stevens, R., Corcho, O. & Gibson, F. 2009, 'Developing ontologies in decentralised settings'. *Nature Precedings*, viewed 28th September 2009, http://precedings.nature.com/documents/3231/version/1.

Garshol, L. M. 2004, 'Metadata? Thesauri? Taxonomies? Topic maps! Making sense of it all', *Journal of Information Science*, vol. 30, no. 4**,** pp. 378-391.

GBRMPA 2005, *Research project ID 133 - long-term monitoring of sea temperatures including at PCQ ports* Australian Government, viewed 20th April 2009, http://www.gbrmpa.gov.au/corp_site/info_services/science_management/research_priorities/database/projects/index.cfm.

GBRMPA 2009, *Coral bleaching forecast and status for the Great Barrier Reef*, Great Barrier Reef Marine Park Authority, viewed 30th June 2009, http://www.gbrmpa.gov.au/corp_site/key_issues/climate_change/management_responses/coral_bleaching_status_2008-09.

GBROOS 2008, *Great Barrier Reef Ocean Observing System*, Integrated Marine Observing System (IMOS), viewed 30th September 2009, http://www.imos.org.au/nodes/great-barrier-reef-observing-system.html.

Gil, Y., Deelman, E., Ellisman, M., Fahringer, T., Fox, G., Gannon, D., Goble, C., Livny, M., Moreau, L. & Myers, J. 2007, 'Examining the challenges of scientific workflows', *IEEE Computer*, vol. 40, no. 12**,** pp. 24-32.

Gleeson, M. W. & Strong, A. E. 1995, 'Applying MCSST to coral reef bleaching', *Advances in Space Research*, vol. 16, no. 10**,** pp. 151-154.

GLEON 2009, *The Global Lake Ecological Observatory Network*, University of Wisconsin Regents, viewed 20th October 2009, http://www.gleonrcn.org/.

Goble, C. 2005, 'Using the Semantic Web for e-Science: inspiration, incubation, irritation', *Proceedings from the 4th International Semantic Web Conference (ISWC 2005)*, Springer, Galway, Ireland, vol 3729, pp. 1-3.

Goble, C. & Bechhofer, S. 2005, 'OntoGrid: a Semantic Grid reference architecture'. *CTWatch Quarterly*, vol. 1, no. 4, viewed 5 january 2006, http://www.ctwatch.org/quarterly/articles/2005/11/ontogrid-a-semantic-grid-reference-architecture/.

Goble, C., Corcho, O., Alper, P. & De Roure, D. 2006, 'e-Science and the Semantic Web: a symbiotic relationship', *Proceedings from the 9th International Conference in Discovery Science (DS 2006)*, Springer, Barcelona, Spain, vol 4265, pp. 1-12.

Goble, C. & De Roure, D. 2002, 'The Semantic Web and Grid Computing', in Kashyap, V. & Shklar, L. (Eds.), *Real World Semantic Web Applications*, IOS Press, pp.

Goble, C. & De Roure, D. 2004, 'The Semantic Grid: myth busting and bridge building', *Proceedings from the 16th European Conference on Artificial Intelligence (ECAI 04)*, IOS Press, Valencia, Spain, vol 110, pp. 1129-1135.

Goble, C. & De Roure, D. 2007, 'myExperiment: social networking for workflow-using e-scientists', *Proceedings of the 2nd workshop on Workflows in support of large-scale science (WORKS 07)*, ACM, Monterey, CA, USA, vol 1, pp. 1-2.

Gomez-Perez, A., Corcho, O. & Fernandez-Lopez, M. 2004, Ontological engineering: with examples from the areas of knowledge management, e-Commerce and the Semantic Web. First edition (advanced information and knowledge processing), Springer, London, England.

Grau, B. C., Horrocks, I., Kazakov, Y. & Sattler, U. 2008, 'Modular reuse of ontologies: theory and practice', *Journal of Artificial Intelligence Research*, vol. 31, no. 1**,** pp. 273-318.

Grau, B. C., Horrocks, I., Motik, B., Parsia, B., Patel-Schneider, P. & Sattler, U. 2008, 'OWL 2: The next step for OWL', *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 6, no. 4**,** pp. 309-322.

Gruber, T. 1993, 'A translation approach to portable ontology specifications', *Knowledge Acquisition,* vol. 5, no. 2**,** pp. 199-220.

Guarino, N. 1997, 'Understanding, building and using ontologies', *International Journal of Human-Computer Studies,* vol. 46, no. 2-3**,** pp. 293-310.

Guarino, N. 1998, 'Formal ontology in information systems', *Proceedings of the 1st International Conference Formal Ontology in Information Systems (FOIS 1998)*, IOS Press, Trento, Italy, vol 46, pp. 337.

Guarino, N., Oberle, D. & Staab, S. 2009, 'What is an *ontology*?', in Staab, S. & Studer, R. (Eds.), *Handbook on Ontologies (International Handbooks on Information Systems)*, 2nd ed, Springer, Berlin/Heidelberg, Germany, pp. 1-20.

Hadzic, M., Wongthongtham, P., Dillon, T. & Chang, E. 2009, 'Ontology design approaches', *Ontology-based multi-agent systems,* Springer, Berlin/Heidelberg, Germany, pp. 75-91.

Hall, W., Roure, D. D. & Shadbolt, N. 2009, 'The evolution of the Web and implications for eResearch ', *Phil. Trans. R. Soc. A,* vol. 367, no. 1890**,** pp. 991-1001.

Hendee, J. & Berkelmans, R. 2003, 'Expert system generated coral bleaching alerts for Myrmidon and Agincourt reefs, Great Barrier Reef, Australia', *Proceedings of the Ninth International Coral Reef Symposium*, Indonesian Institute of Sciences, Bali, Indonesia, vol, pp. 1099-1104.

Hendee, J., Gramer, L., Manzello, D. & Jankulak, M. 2008, 'Ecological forecasting for coral reef ecosystems', *Proceedings of the 11th International Coral Reef Symposium (ICRS 08)*, Ft. Lauderdale, FL, USA, 7-11 July, pp. 148.

Hendler, J. 2003, 'Science and the Semantic Web', *Science,* vol. 299, no. 5606**,** pp. 520-521.

Hendler, J. 2007, 'The dark side of the Semantic Web', *Intelligent Systems, IEEE [see also IEEE Intelligent Systems and Their Applications],* vol. 22, no. 1**,** pp. 2-4.

Hendler, J., Shadbolt, N., Hall, W., Berners-Lee, T. & Weitzner, D. 2008, 'Web science: an interdisciplinary approach to understanding the web', *Commun. ACM,* vol. 51, no. 7**,** pp. 60-69.

Henson, C. A., Neuhaus, H., Sheth, A. P., Thirunarayan, K. & Buyya, R. 2009, 'An ontological representation of time series observations on the Semantic Sensor Web', *1st International Workshop on the Semantic Sensor Web (SemSensWeb 2009)*, CEUR, Crete, Greece, vol 468, pp. 79-94.

Hey, A. J. G. & Trefethen, A. E. 2002, 'The UK e-Science core programme and the Grid', *Future Generation Computer Systems,* vol. 18, no. 8**,** pp. 1017-1031.

Hey, T. & Trefethen, A. E. 2003a, 'The Data Deluge: an e-Science perspective', in Berman F, Fox GC & Hey AJG (Eds.), *Grid Computing - Making the Global Infrastructure a Reality*, John Wiley and Sons Ltd., West Sussex, England, pp. 809-824.

Hey, T. & Trefethen, A. E. 2003b, 'e-Science and its implications', *Philosophical Transactions of the Royal Society*, vol. 361, no. 1809**,** pp. 1809-1825.

Hey, T. & Trefethen, A. E. 2005, 'Cyberinfrastructure for e-Science', *Science*, vol. 308, no. 5723**,** pp. 817-821.

Hitzler, P. & Parsia, B. 2009, 'Ontologies and rules ', in Staab, S. & Studer, R. (Eds.), *Handbook on Ontologies (International Handbooks on Information Systems)*, 2nd ed., Springer, Berlin/Heidelberg, Germany, pp. 111-133.

Hoegh-Guldberg, O. 1999, 'Climate change, coral bleaching and the future of the world's coral reefs', *Marine and Freshwater Research*, vol. 50, no. 8**,** pp. 839-866.

Holmes, G. R. 2008a, 'Developing a scenario-based coral reef ecosystem model to assist management following mass coral mortality events', Doctor of Philosophy, The University of Queensland, Brisbane, Australia.

Holmes, G. R. 2008b, 'Estimating three-dimensional surface areas on coral reefs', *Journal of Experimental Marine Biology and Ecology*, vol. 365, no. 1**,** pp. 67-73.

Horrocks, I. 2002, 'DAML+OIL: a reason-able web ontology language', *Web services, e-Business, and the Semantic Web*, Springer, Berlin / Heidelberg, Germany, pp. 174-174.

Horrocks, I., Parsia, B., Patel-Schneider, P. & Hendler, J. 2005, 'Semantic Web architecture: stack or two towers?', *Proceedings from the 3rd International Workshop Principles and Practice of Semantic Web Reasoning (PPSWR 2005)*, Springer, Dagstuhl Castle, Germany, vol 3703, pp. 37-41.

Horrocks, I., Patel-Schneider, P. F., Boley, H., Tabet, S., Grosof, B. & Dean, M. 2004, *SWRL: a Semantic Web Rule Language - combining OWL and RuleML*, W3C Recommendation viewed 30th July 2009, http://www.w3.org/Submission/SWRL/.

Horrocks, I., Patel-Schneider, P. F. & van Harmelen, F. 2003, 'From SHIQ and RDF to OWL: the making of a web ontology language', *Journal of Web Semantics*, vol. 1, no. 1**,** pp. 7-26.

Huberman, B. & Adamic, L. 1999, 'Internet: growth dynamics of the World-Wide Web', *Nature*, vol. 401, no. 6749**,** pp. 131-131.

Huddlestone-Holmes, C., Gigan, G. & Atkinson, I. 2007, 'Infrastructure for a sensor network on Davies Reef, Great Barrier Reef', *3rd International Conference Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP 07)*, Melbourne, Australia, 3-6 December, IEEE Xplore.

Hughes, T. P., Baird, A. H., Bellwood, D. R., Card, M., Connolly, S. R., Folke, C., Grosberg, R., Hoegh-Guldberg, O., Jackson, J. B. C., Kleypas, J., Lough, J. M., Marshall, P., Nystrom, M., Palumbi, S. R., Pandolfi, J. M., Rosen, B. & Roughgarden, J. 2003, 'Climate change, human impacts, and the resilience of coral reefs', *Science*, vol. 301, no. 5635**,** pp. 929-933.

IMaRS 2009, *Institute for marine remote sensing - millennium coral reef mapping*

College of Marine Science, University of South Florida, viewed 20th April 2009, http://www.imars.usf.edu/MC/imagedb.html.

IMOS 2008, *Integrated Marine Observing System*, viewed June 2008, http://imos.org.au/.

IPCC 2007, 'Climate change 2007: the physical basis', *Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge Univ. Press, Cambridge, U. K.

Jarrar, M. & Meersman, R. 2008, 'Ontology engineering - the DOGMA approach', *Advances in Web Semantics I*, Springer, Berlin/Heidelberg, Germany, pp. 7-34.

Jess 2006, *The rule engine for the Java platform*, viewed January 2009, http://herzberg.ca.sandia.gov/jess/.

Johnston, W. E. 2004, 'Semantic services for Grid-based, large-scale Science', *IEEE Intelligent Systems*, vol. 19, no. 1**,** pp. 34-39.

Jones, R., Hoegh-Guldberg, O. & Larkum, A. 1998, 'Temperature-induced bleaching of corals begins with impairment of the CO2 fixation mechanism in zooxanthellae', *Plant Cell and Environment*, vol. 21, no. 12**,** pp. 1219-1230.

Kininmonth, S., Bainbridge, S., Atkinson, I., Gilla, E., Barrald, L. & Vidaude, R. 2004, 'Sensor networking the Great Barrier Reef', *Spatial Sciences Qld. Journal* vol. Spring 2004, no. 1**,** pp. 34-38.

Klyne, G., Carroll, J. J. & McBride, B. 2004, *Resource Description Framework (RDF): concepts and abstract syntax*, W3C Recommendation, viewed February 2009, http://www.w3.org/TR/rdf-concepts/.

Kno.e.sis.Centre 2008, *Semantic Sensor Web*, Wright State University, Ohio, US, viewed April 2009, http://knoesis.org/research/semsci/application_domain/sem_sensor/.

Köhler, J., Philippi, S., Specht, M. & Rüegg, A. 2006, 'Ontology based text indexing and querying for the semantic web', *Know.-Based Syst.*, vol. 19, no. 8**,** pp. 744-754.

Lacy, L. 2005, OWL: representing information using the Web Ontology Language, Trafford Publishing, Victoria, BC, Canada.

Lassila, O. 1998, 'Web metadata: a matter of semantics', *IEEE Internet Computing*, vol. 2, no. 4**,** pp. 30-37.

Lassila, O. & McGuinness, D. L. 2001, 'The role of frame-based representation on the Semantic Web', *Knowledge Systems Laboratory Technical Report*, Stanford University, Stanford, CA, USA.

Liu, G., Meyer, J. E., Guch, I. C. & Toscano, M. A. 2001, 'NOAA's satellite coral reef bleaching early warning products aimed at local reef sites around the globe', *Reef Encounter*, vol. 30, no. 1**,** pp. 10-13.

Ludäscher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger, E., Jones, M., Lee, E. A., Tao, J. & Zhao, Y. 2006, 'Scientific workflow management and the Kepler system', *Concurrency and Computation: Practice and Experience*, vol. 18, no. 10**,** pp. 1039-1065.

Ludäscher, B., Lin, K., Jaeger-Frank, E., Altintas, S. B. I. & Baru, C. 2007, 'Enhancing the GEON cyberinfrastructure: semantic registration, discovery, mediation, and workflows', *Geosciences Network (GEON) Report*, San Diego, CA, USA.

Manola, F. & Miller, E. 2004, *RDF Primer*, W3C Recommendation, viewed 20 October 2009, http://www.w3.org/TR/rdf-primer/.

Marshall, P. & Schuttenberg, H. 2006, A reef manager's guide to coral bleaching, Great Barrier Reef Marine Park Authority, Townsville, Australia.

Maynard, J., Anthony, K., Marshall, P. & Masiri, I. 2008, 'Major bleaching events can lead to increased thermal tolerance in corals', *Marine Biology*, vol. 155, no. 2**,** pp. 173-182.

Maynard, J. A. 2004, 'Spatial and temporal variation in bleaching susceptibility and the ability of SST variables to describe patterns in bleaching response, 1998 and 2002 event, central GBR.', MSc, James Cook University, Townsville, Australia.

Maynard, J. A., Turner, P. J., Anthony, K. R. N., Baird, A. H., Berkelmans, R., Eakin, C. M., Johnson, J., Marshall, P. A., Packer, G. R., Rea, A. & Willis, B. L. 2008, 'ReefTemp: an interactive monitoring system for coral bleaching using high-resolution SST and improved stress predictors', *Geophys Res Lett.*, vol. 35, no. L05603**,** pp. 1-5.

McGuinness, D. L. 2002, 'Ontologies come of age', in Fensel, D., Hendler, J. A., Lieberman, H. & Wahlster, W. (Eds.), *Spinning the Semantic Web: bringing the World Wide Web to its full potential*, The MIT Press, Cambridge, MA, USA, pp. 171-194.

McGuinness, D. L. 2004, 'Question answering on the semantic Web', *Intelligent Systems, IEEE*, vol. 19, no. 1**,** pp. 82-85.

McGuinness, D. L. & Harmelen, F. v. 2004, *OWL Web Ontology Language Overview* W3C Recommendation viewed February 2009, http://www.w3.org/TR/owl-features/.

McGuinness, D. L. & van Harmelen, F. 2004, *OWL: Web Ontology Language overview* W3C Recommendation viewed February 2009, http://www.w3.org/TR/owl-features/.

Michener, W., Beach, J., Bowers, S., Downey, L., Jones, M., Ludäscher, B., Pennington, D., Rajasekar, A., Romanello, S., Schildhauer, M., Vieglais, D. & Zhang, J. 2005, 'SEEK: data integration and workflow solutions for ecology', *Proceedings from the 2nd International Workshop on Data Integration in the Life Sciences (DILS 2005)*, Springer, San Diego, CA, USA, vol 3615, pp. 321-324.

Mindswap 2007, *Pellet: the open source OWL DL reasoner*, Maryland Information and Network Dynamics Lab Semantic Web Agents Project, viewed July 2007, http://clarkparsia.com/pellet.

MMI 2009, *Marine Metadata Interoperability*, viewed April 2009, http://marinemetadata.org.

Motik, B., Sattler, U. & Studer, R. 2005, 'Query answering for OWL-DL with rules', *Web Semantics: Science, Services and Agents on the World Wide Web* vol. 3, no. 1**,** pp. 41-60.

Myers, T. S., Atkinson, I. & Johnstone, R. 2008, 'Supporting coral reef ecosystems research through modelling re-usable ontologies', *Proceedings from the Knowledge Representation Ontology Workshop (KROW 2008)*, Sydney, Australia, 17 September, ACS, pp. 51-59.

Myers, T. S. & Atkinson, I. M. 2009, 'The Semantic Reef: A hypothesis-based, eco-informatics platform to support automated knowledge discovery for remotely monitored reef systems.', *Proceedings of the 11th International Coral Reef Symposium (ICRS 08)*, Ft. Lauderdale, FL, USA, 7-11 July, pp. 154.

Myers, T. S., Atkinson, I. M. & Johnstone, R. 2009, 'Supporting coral reef ecosystems research through modelling a re-usable ontology framework', *Journal of Applied Artificial Intelligence*, vol. 90, no. 24**,** pp. (in press).

Myers, T. S., Atkinson, I. M. & Maynard, J. 2007, 'The Semantic Reef: An eco-informatics approach for modelling coral bleaching within the Great Barrier Reef', *Environmental Research Event (ERE 07)* Cairns, Australia, 1 December, Environmental Research Event Organising Committee.

Nambiar, U., Ludaescher, B., Lin, K. & Baru, C. 2006, 'The GEON portal: accelerating knowledge discovery in the geosciences', *Proceedings of the 8th annual ACM International Workshop on Web Information and Data Management (WIDM 06)*, Arlington, VA, USA, 10-12 November, ACM, pp. 83 - 90.

NASA 2009, *Semantic Web for Earth and Environmental Terminology (SWEET)* Jet Propulsion Laboratory, California Institute of Technology, viewed July 2009, http://sweet.jpl.nasa.gov/ontology/.

Neiswender, C. 2009, *"What is a controlled vocabulary?" in the MMI Guides: navigating the world of marine metadata*, viewed April 2009, http://marinemetadata.org/guides/vocabs/vocdef.

NEON 2008, *The National Ecological Observatory Network*, NEON, Inc. , viewed April 2009, http://www.neoninc.org/.

NEPTUNE 2009, *North-East Pacific Time-series Undersea Networked Experiments - NEPTUNE Cananda: transforming Ocean Science*, Ocean Networks Canada (ONC), University of Victoria, viewed April 2009, http://www.neptunecanada.ca/.

NeSC 2009, *National e-Science Centre*, viewed February 2009, http://www.nesc.ac.uk/nesc/define.html.

NOAA-ICON/CREWS 2008, *Integrated Coral Observing Network/Coral Reef Early Warning System*, National Oceanic and Atmospheric Administration, viewed May 2008, http://www.coral.noaa.gov/crews/.

NOAA 2009a, *Coral Reef Watch*, viewed May 2009, http://coralreefwatch.noaa.gov/satellite/index.html.

NOAA 2009b, *Coral Reef Watch - methodology and description*, viewed May 2009, http://coralreefwatch.noaa.gov/satellite/methodology/methodology.html.

NOAA/CHAMP 2006, *NOAA's Coral Health and Monitoring Program (CHAMP)*, viewed April 2009, http://www.coral.noaa.gov/index.shtml.

Noy, N. F. 2004, 'Semantic integration: a survey of ontology-based approaches', *SIGMOD Rec.*, vol. 33, no. 4**,** pp. 65-70.

Noy, N. F. & McGuinness, D. L. 2001, 'Ontology development 101: a guide to creating your first ontology'. *Knowledge Systems Laboratory*, viewed 20th February, 2006, http://ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness.pdf.

O'Connor, M., Shankar, R. & Das, A. 2006, 'An ontology-driven mediator for querying time-oriented biomedical data', *Proceedings from the 19th IEEE International Symposium on Computer-Based Medical Systems (CBMS 06)*, IEEE, Salt Lake City, UT, USA, vol 5, pp. 264-269.

O'Connor, M. J., Knublauch, H., Tu, S. W., Grossof, B., Dean, M., Grosso, W. E. & Musen, M. A. 2005, 'Supporting rule system interoperability on the Semantic Web with SWRL', *Fourth International Semantic Web Conference (ISWC-2005)*, Springer Berlin / Heidelberg, Galway, Ireland, vol 3729, pp. 974-986.

O'Connor, M. J., Nyulas, C. I., Shankar, R. D., Das, A. K. & Musen, M. A. 2008, 'The SWRLAPI: a development environment for working with SWRL rules', *Fifth International Workshop on OWL: Experiences and Directions (OWLED 08), held with 7th International Semantic Web Conference*, Karlsruhe, Germany, 26-27 October.

O'Connor, M. J., Tu, S. W., Nyulas, C. I., Das, A. K. & Musen, M. A. 2007, 'Querying the Semantic Web with SWRL', *The International RuleML Symposium on Rule Interchange and Applications (RuleML2007)*, Springer Verlag, Orlando, FL, USA, vol 4824, pp. 155-159.

O'Hara, K. & Hall, W. 2009, 'Semantic Web', in Bates, M. J., Maack, M. N. & Drake, M. (Eds.), *Encyclopedia of Library and Information Science*, 2nd ed., Taylor & Francis, London, England, pp. (in press).

Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., Carver, T., Glover, K., Pocock, M. R., Wipat, A. & Li, P. 2004, 'Taverna: a tool for the composition and enactment of bioinformatics workflows', *Bioinformatics*, vol. 20, no. 17**,** pp. 3045-3054.

Olsson, P., Folke, C. & Hughes, T. P. 2008, 'Navigating the transition to ecosystem-based management of the Great Barrier Reef, Australia', *Proc Natl Acad Sci USA*, vol. 105, no. 28**,** pp. 9489-9494.

OntoGrid 2007, 'Final report - systematic metadata management for applications that use the Grid', *OntoGrid: Paving the way for Knowledgeable Grid Services and Systems IST FP6-511513* Manchester, UK.

Page, R. D. M. 2006, 'Taxonomic names, metadata, and the Semantic Web', *Biodiversity Informatics*, vol. 3, no. 1**,** pp. 1-15.

Pordes, R., Petravick, D., Kramer, B., Olson, D., Livny, M., Roy, A., Avery, P., Blackburn, K., Wenaus, T., Wurthwein, F., Foster, I., Gardner, R., Wilde, M., Blatecky, A., McGee, J. &

Quick, R. 2008, 'The Open Science Grid status and architecture', *Journal of Physics: Conference Series*, vol. 119, no. 5**,** pp. 052028.

Powers, S. 2003, Practical RDF, O'Reilly & Associates, Inc., Sebastopol, CA, USA.

Protégé 2009, *The ontology editor and knowledge acquisition system*, Stanford University, viewed April 2009, http://protege.stanford.edu/.

Prud'hommeaux, E. & Seaborne, A. 2008, *SPARQL query language for RDF*, W3C, viewed February 2009, http://www.w3.org/TR/rdf-sparql-query/.

RacerPRO 2008, *Renamed ABox and concept expression reasoner*, Racer Systems GmbH and Co, viewed June 2008, http://www.racer-systems.com/.

Rajasegarar, S., Gubbi, J., Bondarenko, O., Kininmonth, S., Marusic, S., Bainbridge, S., Atkinson, I. & Palaniswami, M. 2008, 'Sensor network implementation challenges in the Great Barrier Reef marine environment', *ICT Mobile and Wireless Communications Summit (ICT-MobileSummit 2008)*, Stockholm, Sweden, 10-12 June.

Rector, A. 2003, 'Modularisation of domain ontologies implemented in Description Logics and related formalisms including OWL', *Proceedings of the 2nd International Conference on Knowledge Capture (K-CAP 03)*, Sanibel Island, FL, USA, ACM Press, pp. 121-128.

Rector, A., Drummond, N., Horridge, M., Rogers, J., Knublauch, H., Stevens, R., Wang, H. & Wroe, C. 2004, 'OWL pizzas: practical experience of teaching OWL-DL: common errors and common patterns', *Proceedings of the European Conference on Knowledge Acquistion (EKAW 2004)*, Springer-Verlag, Northampton, England, vol 3257, pp. 63-81.

Sanchez-Gestido, M., Blanco-Abruna, L., Perez-Hernandez, M. S., Gonzalez-Cabrero, R., Gomez-Perez, A. & Corcho, O. 2006, 'Complex data-intensive systems and Semantic Grid: applications in satellite missions', *the 2nd IEEE International Conference on e-Science and Grid Computing (e-Science)*, IEEE Computer Society, Amsterdam, Netherlands, vol, pp. 158-158.

Schuhmacher, H. & Zibrowius, H. 1985, 'What is hermatypic?', *Coral Reefs*, vol. 4, no. 1**,** pp. 1-9.

SEEK 2009, *Science Environment for Ecological Knowledge*, viewed March 2009, http://seek.ecoinformatics.org/.

Shadbolt, N., Berners-Lee, T. & Hall, W. 2006, 'The Semantic Web revisited', *IEEE Intelligent Systems*, vol. 21, no. 3**,** pp. 96-101.

Sheth, A. 2008, 'Semantic Sensor Web', Invited talk, presented at presented at the ARC Research Network on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP 08), Melbourne, Australia, 1 August, Available online at http://knoesis.org/research/semsci/application_domain/sem_sensor/.

Sheth, A., Henson, C. & Sahoo, S. S. 2008, 'Semantic Sensor Web', *IEEE Internet Comput*, vol. 12, no. 4**,** pp. 78-83.

SIOC 2009, *Semantically-Interlinked Online Communities Project*, viewed March 2009, http://sioc-project.org/.

Smith, M. K., Welty, C. & McGuinness, D. L. 2004, *OWL Web Ontology Language guide*, W3C Recommendation viewed February 2009, http://www.w3.org/TR/owl-guide/.

Spalding, M., Ravilious, C. & Green, E. 2001, World atlas of coral reefs, University of California Press, London, England.

SQWRL 2008, *SQWRL*, viewed July 2009, http://protege.cim3.net/cgi-bin/wiki.pl?SQWRL.

Stocks, K. I., Neiswender, C., Isenor, A. W., Graybeal, J., Galbraith, N., Montgomery, E. T., Alexander, P., Watson, S., Bermudez, L., Gale, A. & Hogrefe, K. 2009, *The MMI guides: navigating the world of marine metadata*, viewed April 2009, http://marinemetadata.org/guides.

Strong, A. E., Barrientos, C. S., Duda, C. & Sapper, J. 1997, 'Improved satellite techniques for monitoring coral reef bleaching', *Proceedings of the 8th International Coral Reef Symposium*, Panama City, vol 2, pp. 1495–1498.

Studer, R., Benjamins, R. V. & Fensel, D. 1998, 'Knowledge engineering: principles and methods', *Data Knowl. Eng.*, vol. 25, no. 1-2**,** pp. 161-197.

Sweatman, H., Abdo, D., Burgess, S., Cheal, A., Coleman, G., Delean, S., Emslie, M., Miller, I., Osborne, K., Oxley, W., Page, C. & Thompson, A. 2003, 'Long-term monitoring of the Great Barrier Reef - status report number 6', *Reefs of the GBR - status and trends*, AIMS, Townsville, Australia.

Szalay, A. & Gray, J. 2006, '2020 computing: science in an exponential world', *Nature*, vol. 440, no. 7083**,** pp. 413-414.

Taylor, I., Deelman, E., Gannon, D. & Shields, M. (Eds.) 2006, *Workflows for e-Science: scientific workflows for Grids*, Springer, Berlin/Heidelberg, Germany.

Taylor, I., Shields, M., Wang, I. & Harrison, A. 2007, 'The Triana workflow environment: architecture and applications', *Workflows for e-Science*, Springer, London, England, pp. 320-339.

Taylor, J. A., Zic, J. & Morrissey, J. 2008, 'Building CSIRO e-Research capabilities', *eResearch Australasia 2008*, Melbourne, Australia 28 September-3 October, Australian Department of Innovation, Industry, Science and Research.

Taylor, K. R., Essex, J. W., Frey, J. G., Mills, H. R., Hughes, G. & Zaluska, E. J. 2006, 'The Semantic Grid and chemistry: experiences with CombeChem', *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 4, no. 2**,** pp. 84-101.

Thorpe, A. 2009, 'Environmental eScience', *Phil. Trans. R. Soc. A*, vol. 367, no. 1890**,** pp. 801-802.

Tuecke, S., Czajkowski, K., Foster, I., Frey, J., Graham, S., Kesselman, C., Maquire, T., Sandholm, T., Snelling, D., Vanderbilt, P., 2003, 'Open Grid Services Infrastructure (OGSI) version 1.0', *Open Grid Service Infrastructure WG*, Global Grid Forum (GGF), Chicago, IL, USA.

Unicode 1991-2009, *The Unicode Consortium*, viewed June 2008, http://www.unicode.org/.

Uschold, M. 1996, 'Building ontologies: towards a unified methodology', *16th Annual Conference of the British Computer Society Specialist Group on Expert Systems*, Cambridge, United Kingdom, 16-18 December.

Uschold, M. 2003, 'Where are the semantics in the Semantic Web?', *AI Mag.*, vol. 24, no. 3**,** pp. 25-36.

Uschold, M. & Gruninger, M. 1996, 'Ontologies: principles, methods and applications', *Knowledge Engineering Review*, vol. 11, no. 1**,** pp. 93 - 136.

Uschold, M. & King, M. 1995, 'Towards a methodology for building ontologies', *Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing, International Joint Conference on Artificial Intelligence. 14th International Joint Conference on Artificial Intelligence (IJCAI 1995)*, Montreal, Canada, 20-25 August, Morgan Kaufmann.

van Harmelen, F., Horrocks, I. & Patel-Schneider, P. F. 2001, *A model-theoretic semantics for DAML+OIL (March 2001)*, W3C, viewed February 2009, http://www.w3.org/TR/daml+oil-model.

van Heijst, G., Schreiber, A. T. & Wielinga, B. J. 1997, 'Using explicit ontologies in KBS development', *Int. J. Hum.-Comput. Stud.*, vol. 46, no. 2-3**,** pp. 183-292.

W3C 2004a, *Architecture of the World Wide Web, Volume One*, viewed February 2009, http://www.w3.org/TR/webarch/.

W3C 2004b, *World Wide Web Consortium issues RDF and OWL recommendations*, W3C, viewed April 2005, http://www.w3.org/2004/01/sws-pressrelease.

W3C 2007, 'Emerging Web Technologies', Presentation, presented at Worldwide Web Consortium (W3C), http://www.w3.org/2007/Talks/0130-sb-W3CTechSemWeb.

W3C 2009a, *Design issues - architecture and philosophical points*, World Wide Web Consortium, viewed March 2009, http://www.w3.org/DesignIssues/Overview.html.

W3C 2009b, *The Linking Open Data Project*, viewed March 2009, http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData.

W3C 2009c, *Semantic Web activity*, Worldwide Web Consortium (W3C), viewed January 2009, http://www.w3.org/2001/sw/.

W3C 2009d, *Semantic Web case studies and use cases* Worldwide Web Consortium (W3C), viewed January 2009, http://www.w3.org/2001/sw/sweo/public/UseCases/.

Wache, H., Vogele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H. & Hubner, S. 2001, 'Ontology-based integration of information - a survey of existing approaches', *17th International Joint Conference on Artificial Intelligence (IJCAI 01) Workshop: Ontologies and Information Sharing*, Seattle, WA, USA, 4-10 August, American Association for Artificial Intelligence, pp. 108-117.

Waldrop, M. 2008, 'Science 2.0 - is open access science the future? Is posting raw results online, for all to see, a great tool or a great risk?'. *Scientific American*, issue May 2008, viewed May 2008, http://www.sciam.com/article.cfm?id=science-2-point-0&print=true.

Wolstencroft, K., Brass, A., Horrocks, I., Lord, P., Sattler, U., Turi, D. & Stevens, R. 2005, 'A little Semantic Web goes a long way in biology', *Proceedings from the 4th International Semantic Web Conference (ISWC 2005)*, Springer, Galway, Ireland, vol 3729, pp. 786-800.

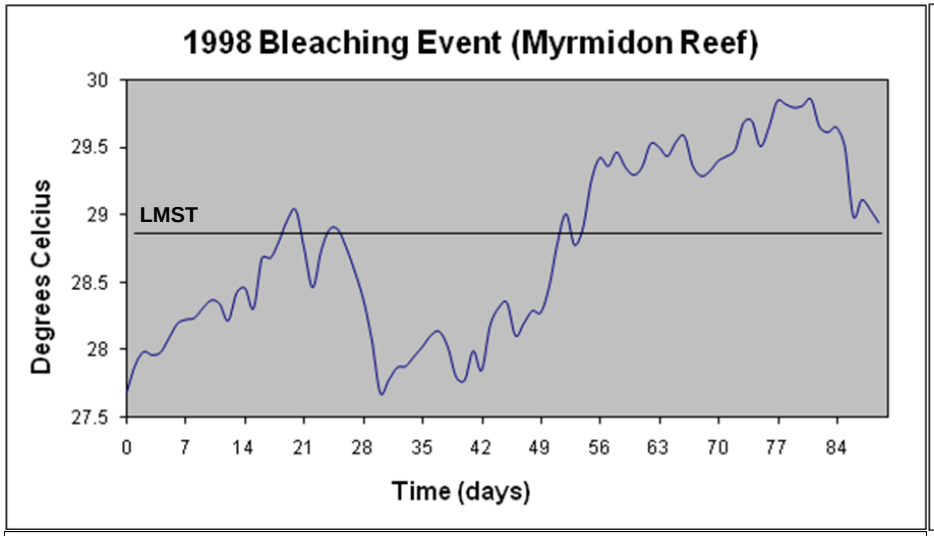
Wright, R., Sánchez-Gestido, M., Gómez-Pérez, A., Pérez-Hernández, M., González-Cabero, R. & Corcho, O. 2008, 'A semantic data grid for satellite mission quality analysis', *The Semantic Web - ISWC 2008*, Springer Berlin / Heidelberg, pp. 818-832.

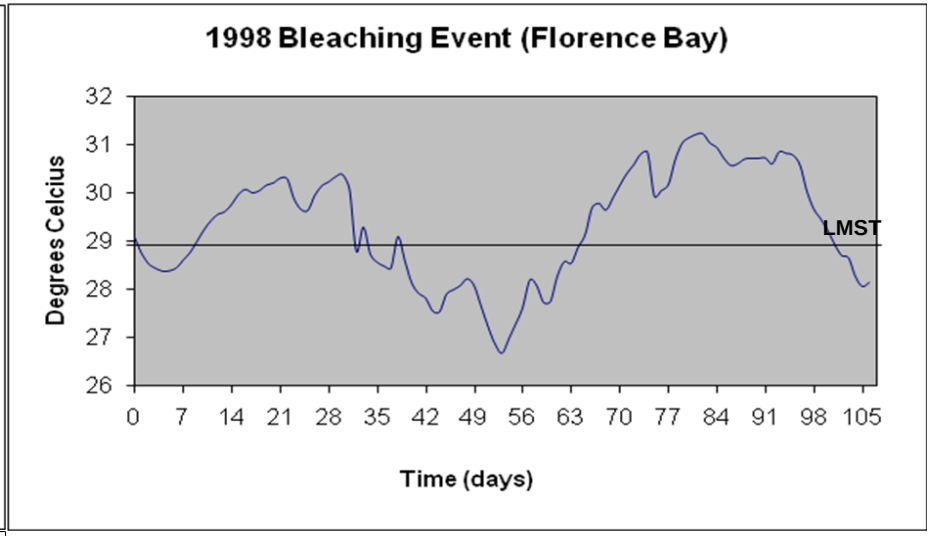
Appendix A-Comparative Analysis of Eco-informatic Systems

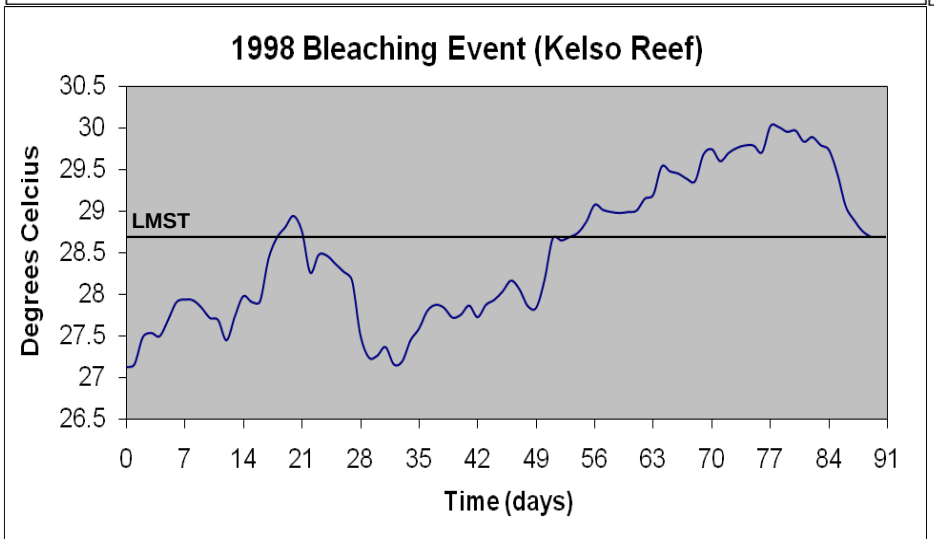
Projects / Characteristic	Semantic Reef	SEEK	Semantic Sensor Web	ICON/ CREWS	ONTOGRID Quarc	Health-e-Waterways
Observational data model	yes +	no -	no -	yes +	no -	yes +
Demands quality assured data	no +	yes -	yes -	yes -	yes -	yes -
Set data sources	no +	no +	yes -	yes -	yes -	yes -
Open data sources	yes +	yes +	no	no	no	no
Data Silo's only	no +	no +	yes -	yes -	yes -	yes -
Sensed near-time or real-time data	yes +	no -	yes +	yes +	no -	yes +
RDF Triplestore	yes +	no -	yes +	no -	yes +	yes +
OWL	yes +	yes +	yes +	no -	no -	no -
Reasoning with DL	yes +	yes - for a purpose	yes - for a purpose	no -	no -	no -
Inference rules	yes +	no -	yes +	no -	no -	no -
Semantic query with SPARQL	yes +	no -	yes +	no -	yes +	yes
Grid computing	yes +	yes +	no -	no -	yes +	no -
Workflows	yes +	yes +	no -	no -	yes +	yes +
Query systems	yes +	yes +	yes +	yes +	yes +	yes +
Hypothesis support	yes +	no -	yes +	no -	no -	no -
Support or use of visualization	yes +	yes +	yes +	yes +	yes +	yes +
Explicit purpose	no + proposition driven	yes - Capture, organize and search analytical processes and data	yes - but can pose other rules	yes -	yes -	yes - Report cards
Scalable to general purpose	yes +	yes +	no -	no -	no -	no -
Web portal	no -	no -	yes - limited	yes +	no -	yes +
Marine Science	yes	yes - holistic ecological outlook	no	yes	no	no
Not Marine Science	no	no	yes - urban information and alert system	no	yes - Earth observation system and imagery	yes - Hydrology-movement consumption quality

Architectural comparison of eco-informatics data integration systems.

Appendix B–1998 Summer SST

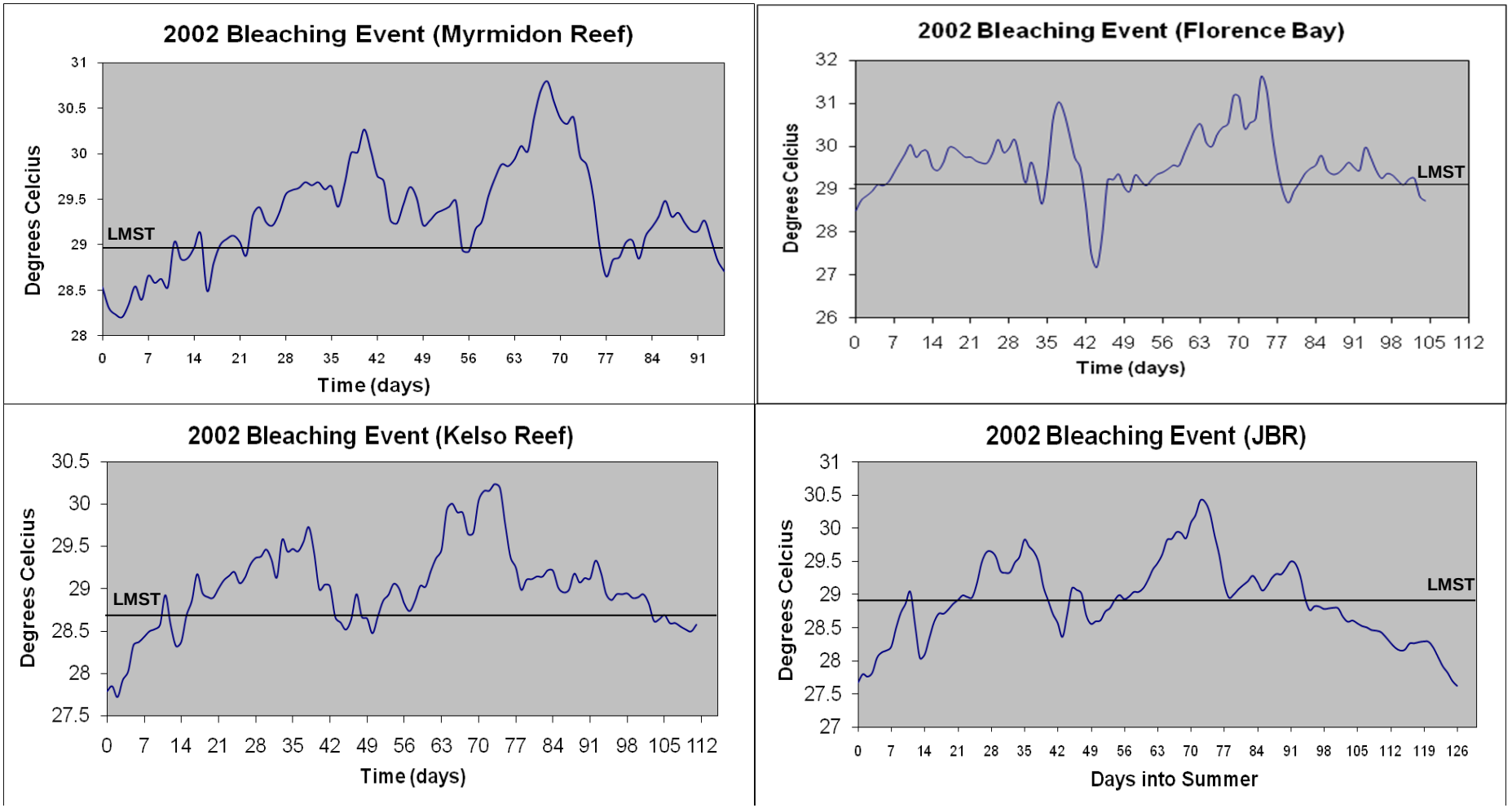






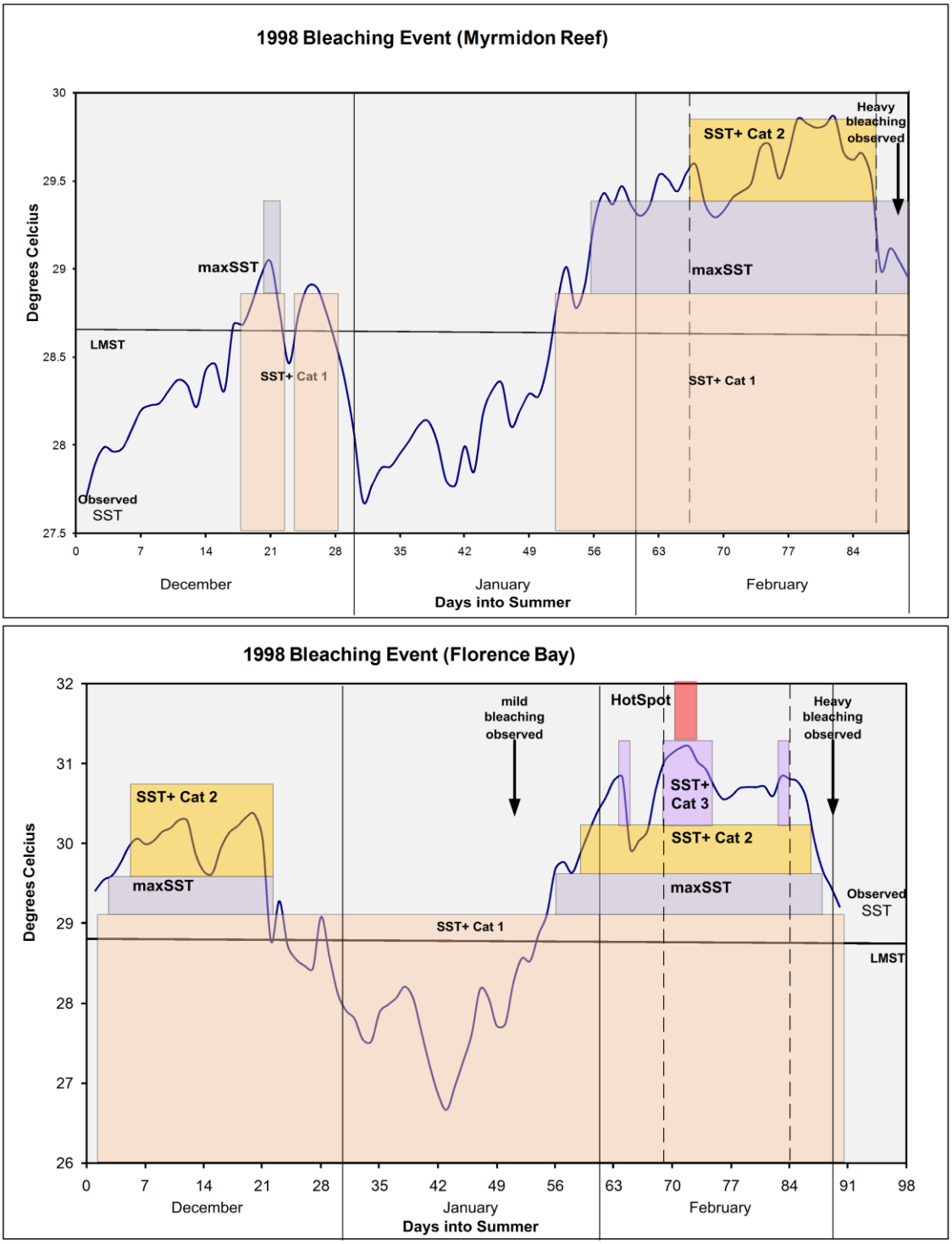
1998 summer SST+ at Myrmidon Reef, Florence Bay and Kelso Reef. 1998 SST data not available for John Brewer Reef (GBRMPA 2005).

Appendix C–2002 Summer SST

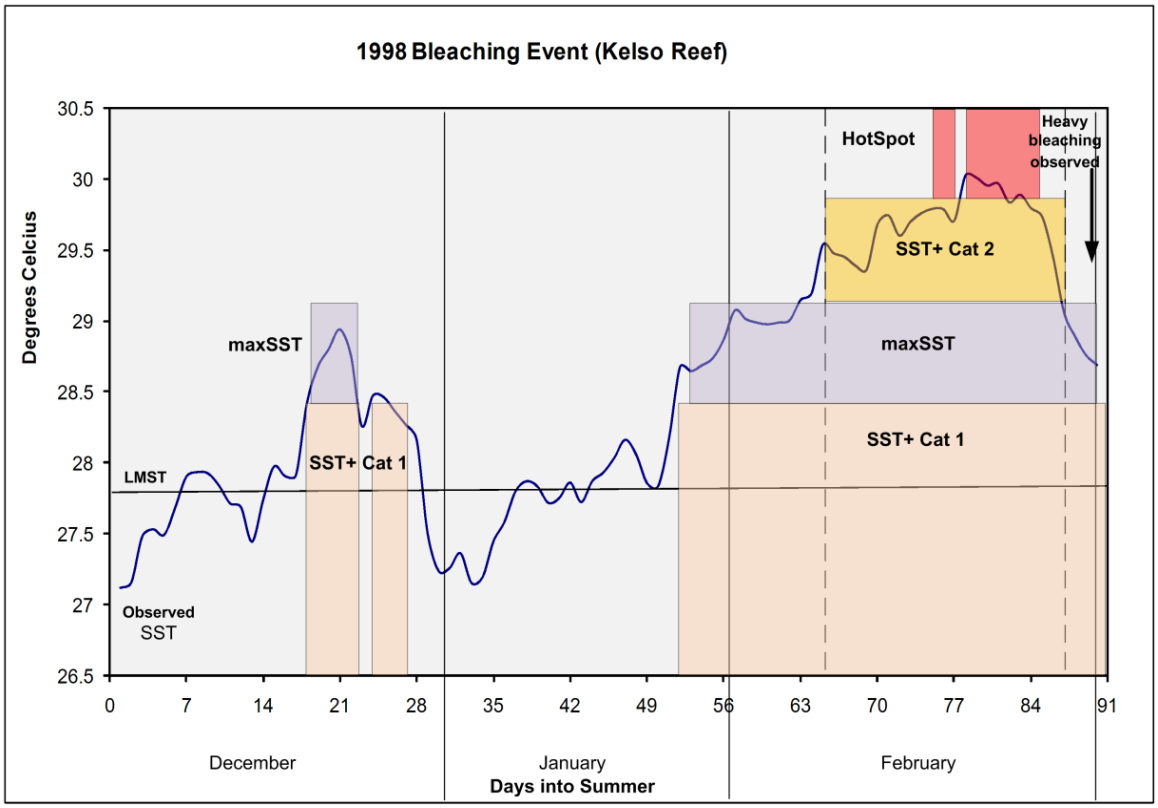


2002 summer SST+ at Myrmidon Reef, Florence Bay, Kelso Reef and John Brewer Reef (GBRMPA 2005).

Appendix D–1998 Results-SST Anomaly Indices

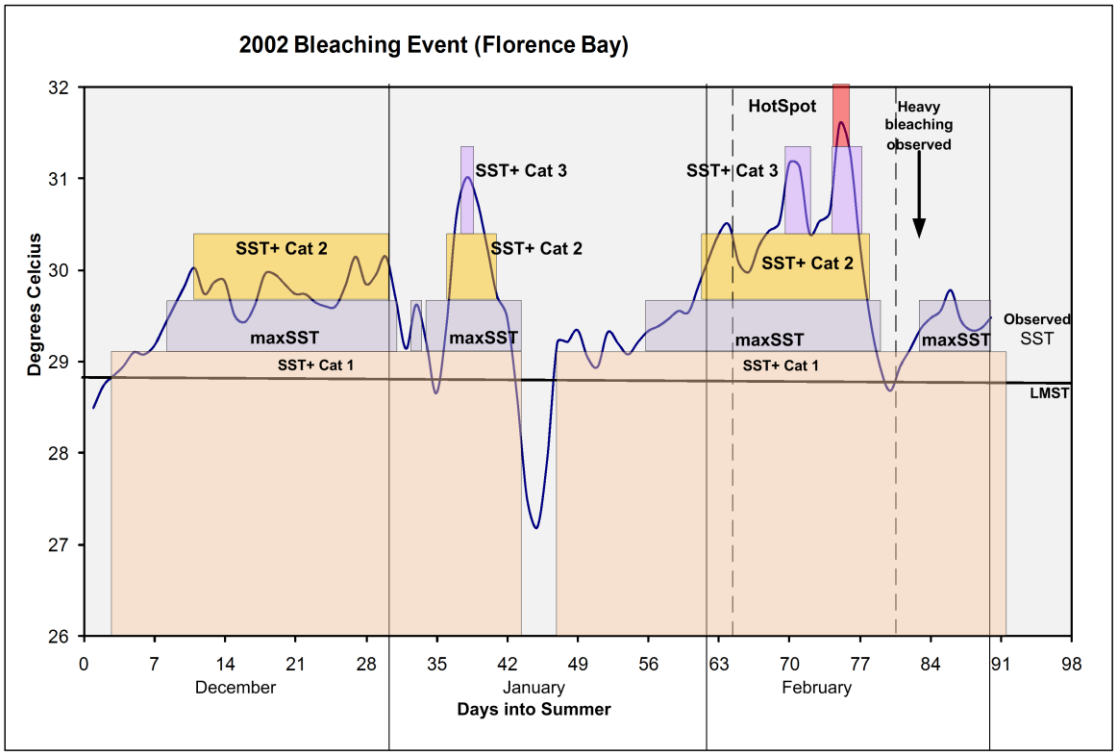


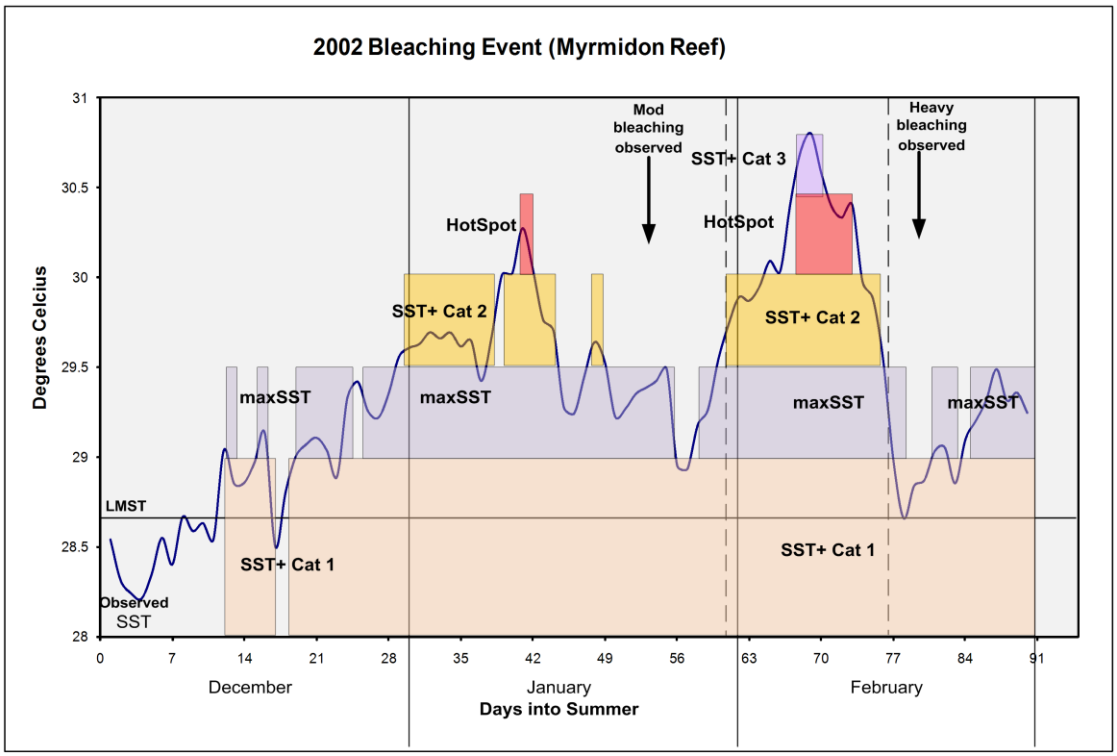
The 1998 summer observed SST for Myrmidon Reef and Florence Bay (GBRMPA 2005). The rectangle overlays are regions that inferred a high risk of coral bleaching.



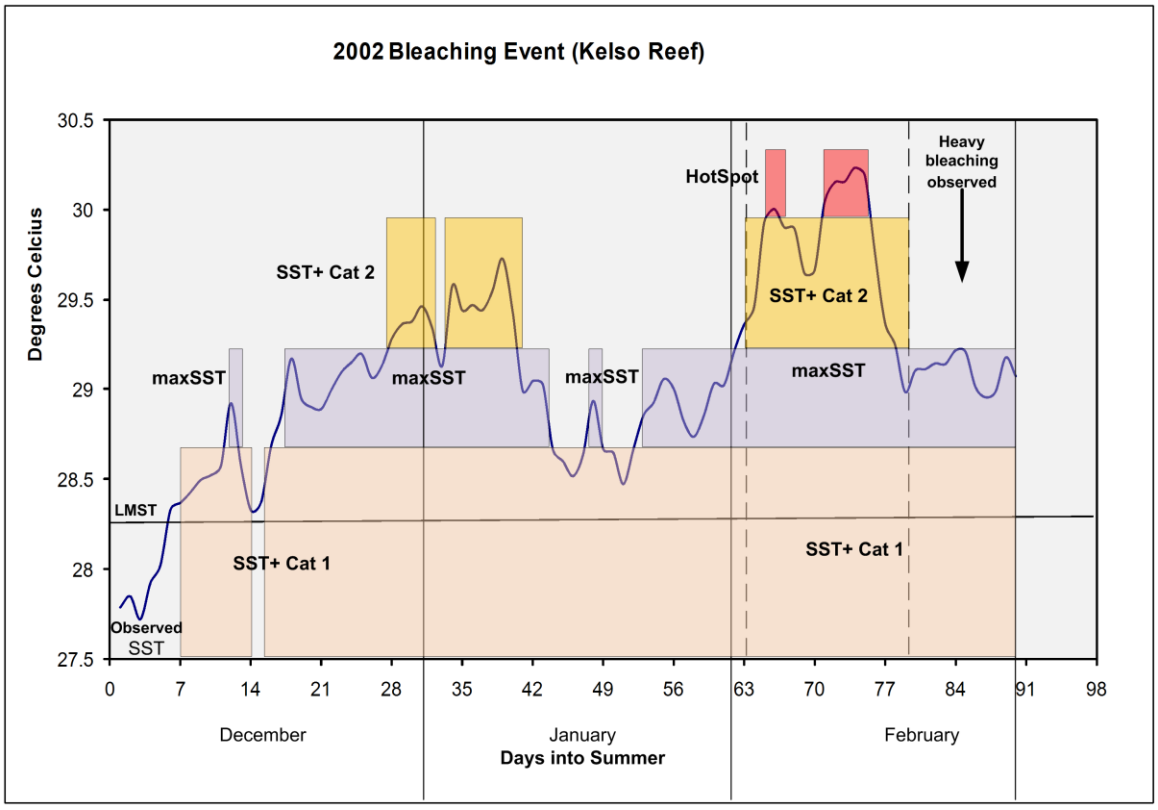
The 1998 observed summer SST for Kelso Reef (GBRMPA 2005). The rectangle overlays are regions that inferred a high risk of coral bleaching.

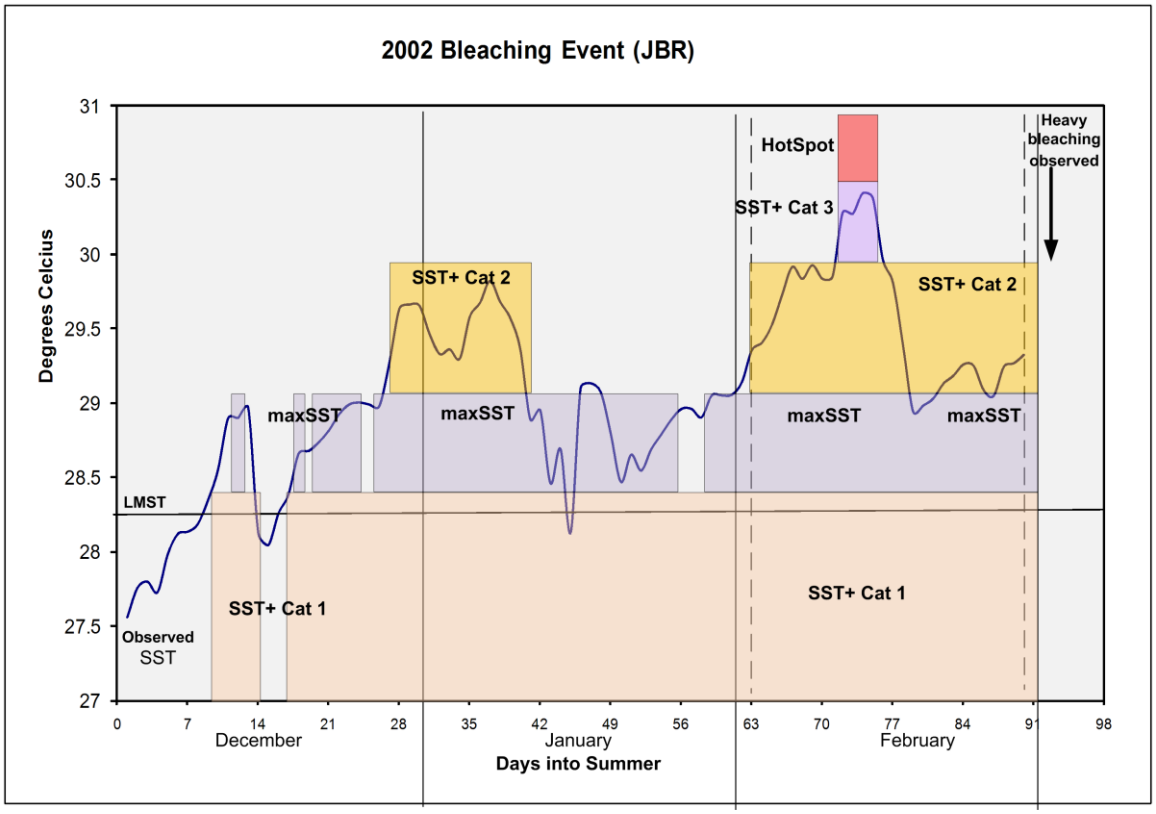
Appendix E–2002 Results-SST Anomaly Indices





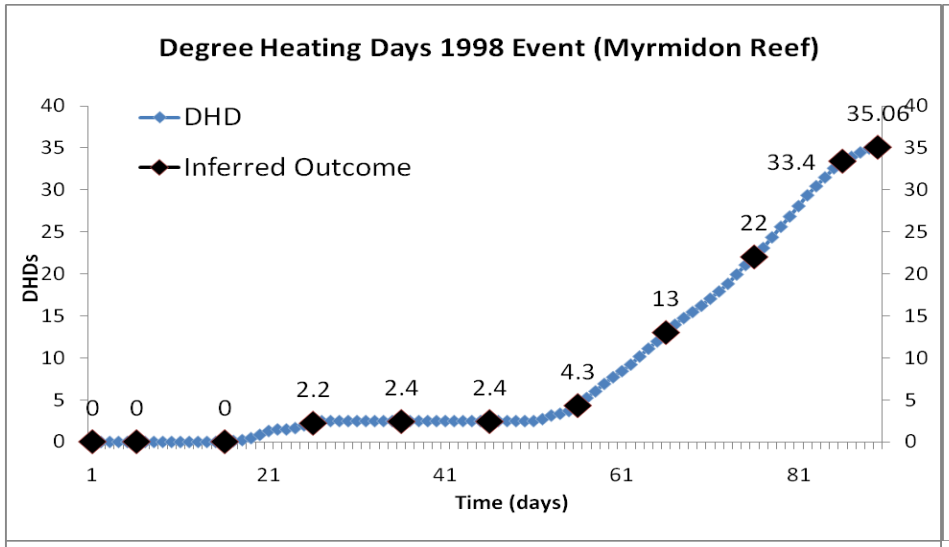
The 2002 observed summer SST for Florence Bay and Myrmidon Reef (GBRMPA 2005). The rectangle overlays are regions that inferred a high risk of coral bleaching.

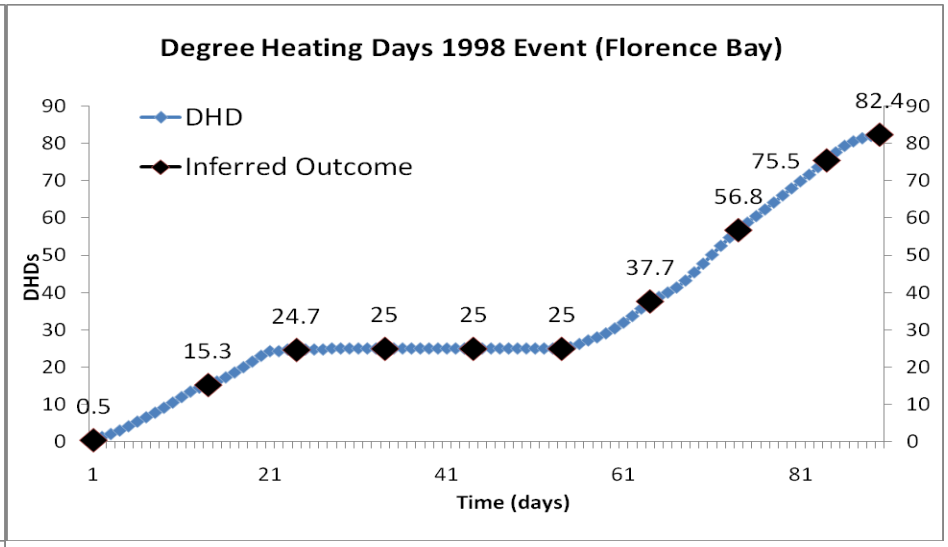


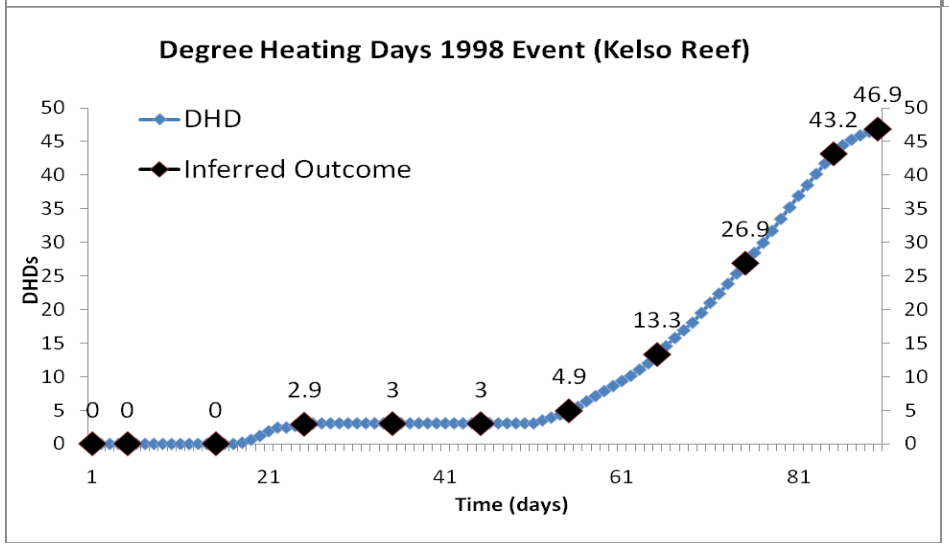


The 2002 observed summer SST for Kelso Reef and John Brewer Reef (GBRMPA 2005). The rectangle overlays are regions that inferred a high risk of coral bleaching.

Appendix F-1998 Results-Summer DHDs

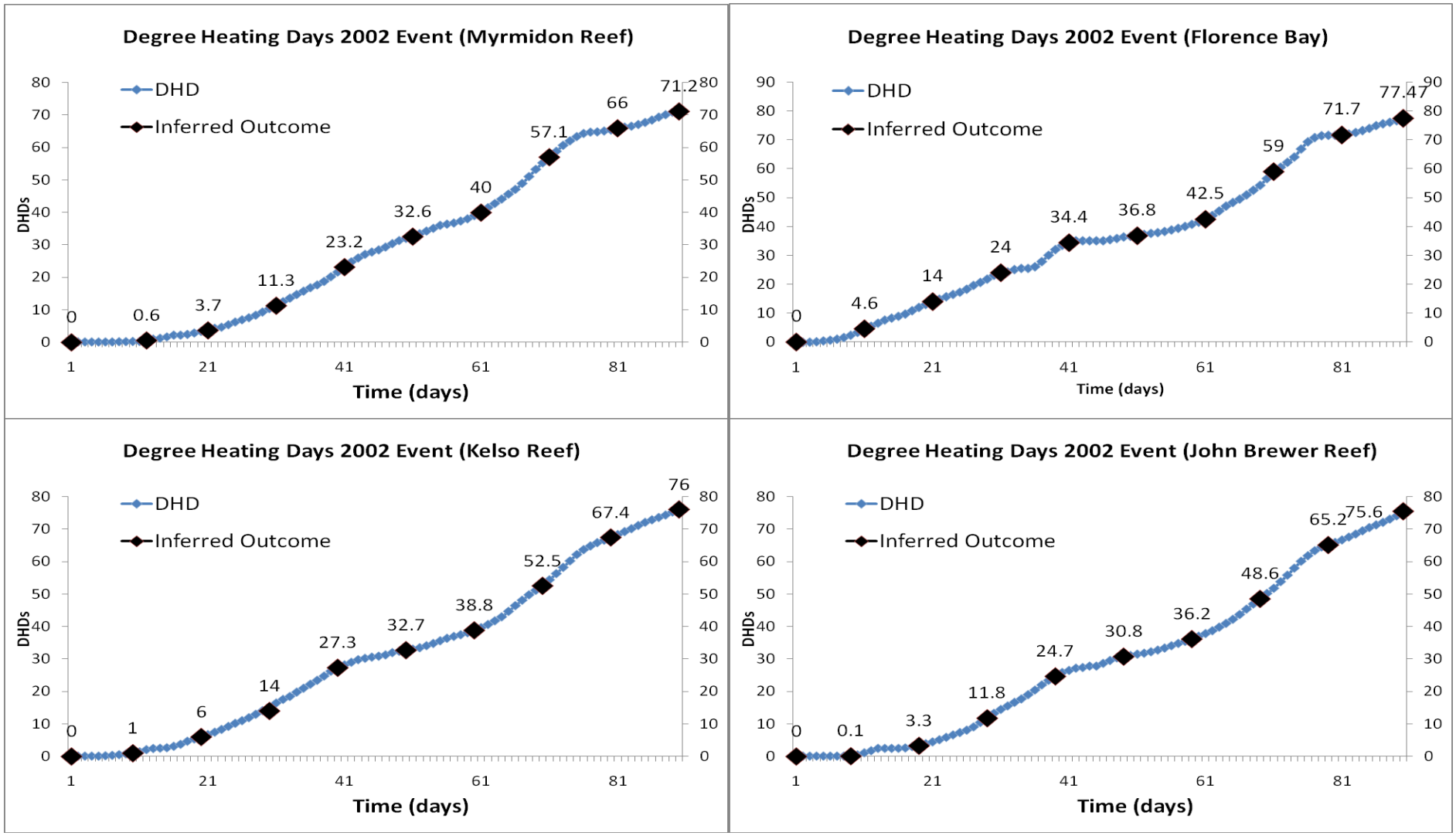






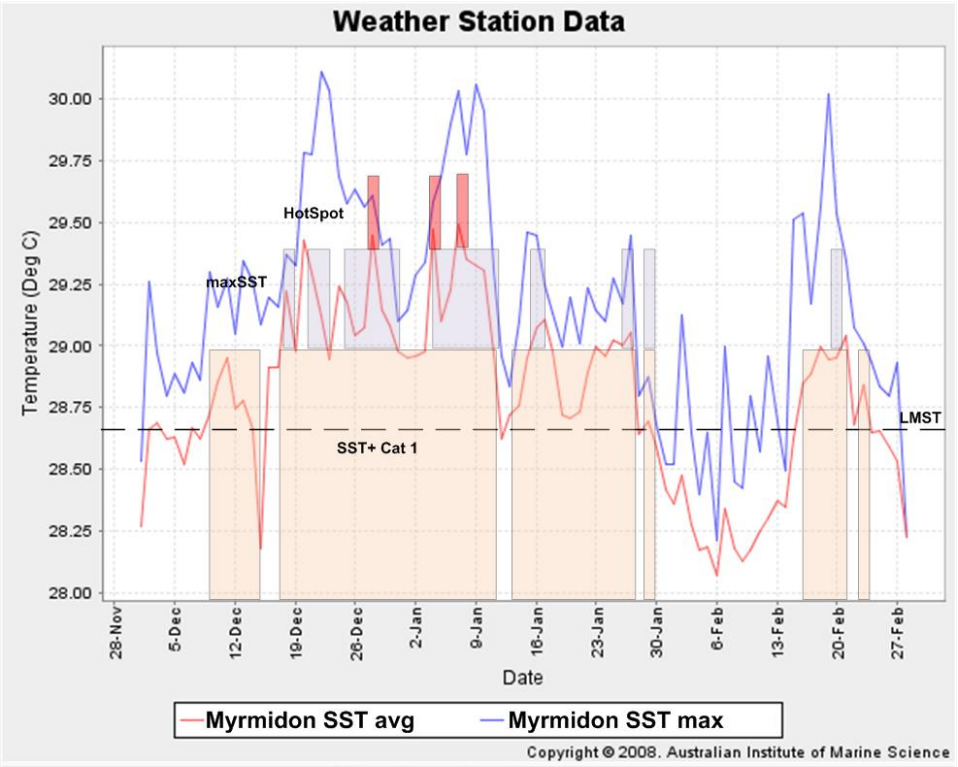
1998 summer DHD (blue) and the outcome from the KB queries (Black) for Myrmidon Reef, Florence Bay and Kelso Reef.

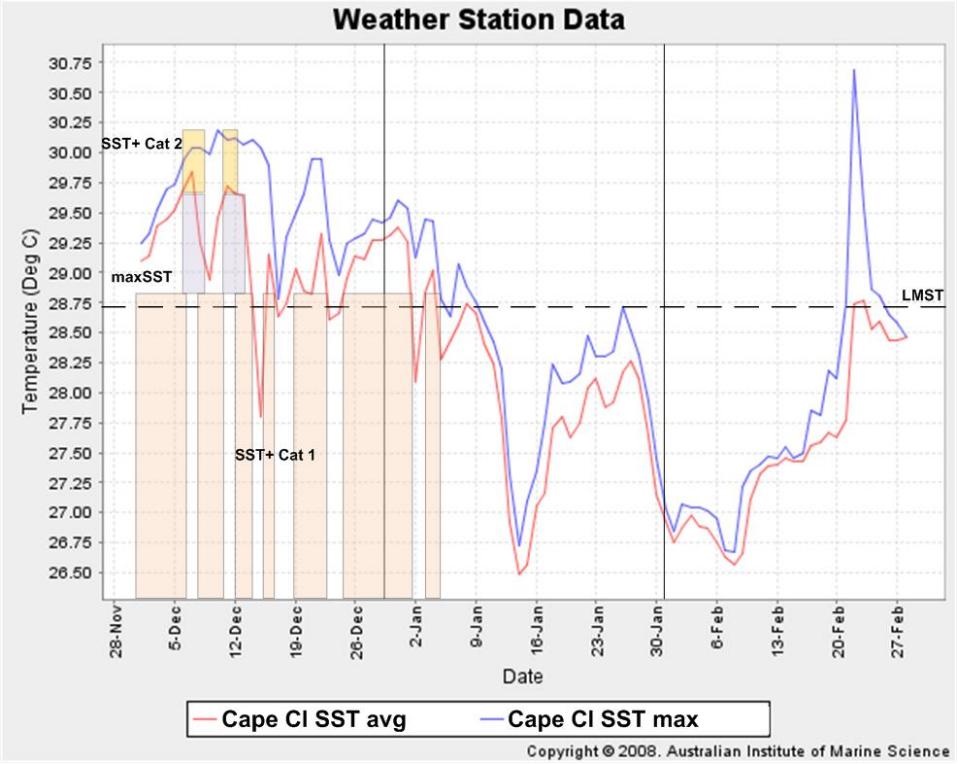
Appendix G-2002 Results-Summer DHDs



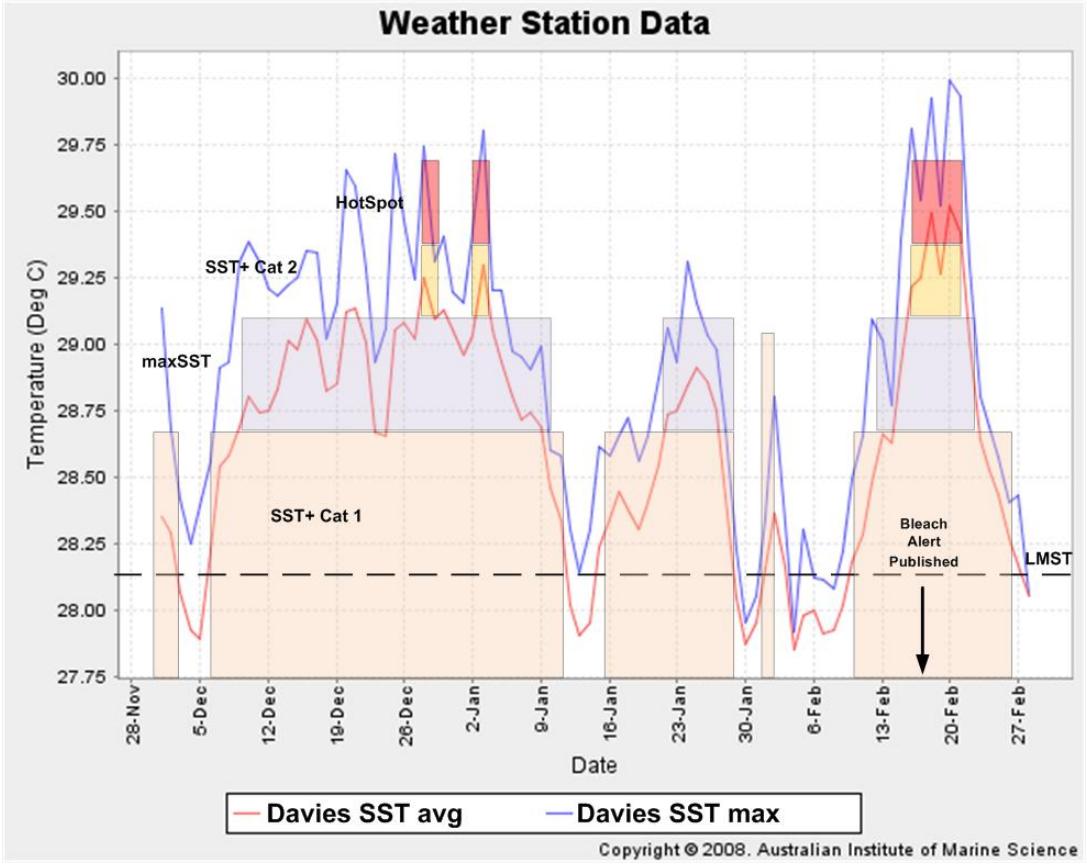
2002 summer DHD (blue) and the outcome from the KB queries (Black) for Myrmidon Reef, Florence Bay, Kelso Reef and John Brewer Reef.

Appendix H-Example 1 SST Indices





The 2009 summer for Myrmidon Reef and Cleveland Bay depicts the last summer temperature outcome where no bleaching occurred.



The 2009 summer for Davies Reef – high temperatures in mid February invoke a bleach alert.

Appendix I–SWRL Inference Rules code

Rule 1 – 5 atoms

```
Coral_Reef:Coral_Reef(?x)  ∧
Coral_Reef:hasDailyAverageSSTof(?x, ?meanTemp)  ∧
Coral_Reef:hasAverageLongTermSeaSurfaceTemperatureOf(?x, ?L
swrlb:greaterThanOrEqual(?meanTemp, ?LMST)
→ SST_PLUS(?x)
```

Rule 2 – 9 atoms

```
Coral_Reef:Coral_Reef(?x)  ∧
Coral_Reef:hasDailyAverageSSTof(?x, ?meanTemp)  ∧
Coral_Reef:hasLongtermMeanMAXSummerSSTOf(?x, ?LongMax)  ∧
swrlb:greaterThan(?meanTemp, ?LongMax)  ∧
Coral_Reef:hasMAXMonthlyMeanSSTOf(?x, ?MMM)  ∧
swrlb:greaterThan(?meanTemp, ?MMM)  ∧
Coral_Reef:hasAverageLongTermSeaSurfaceTemperatureOf(?x, ?LMST)
swrlb:greaterThan(?meanTemp, ?LMST)
→ All_Indices(?x)
```

Rule 3 – 16 atoms

```
Coral_Reef:Coral_Reef(?x)  ∧
Coral_Reef:hasDateTimeOf(?x, ?dateTime)  ∧
Coral_Reef:hasDailyAverageSSTof(?x, ?meanTemp)  ∧
Coral_Reef:hasAverageLongTermSeaSurfaceTemperatureOf(?x, ?LMST) ∧
swrlb:greaterThanOrEqual(?meanTemp, ?LMST)  ∧
Coral_Reef:hasTurbidityLevelOf(?x, ?turb)  ∧
swrlb:greaterThanOrEqual(?turb, 80)  ∧
Coral_Reef:hasLightQuantaOf(?x, ?par)  ∧
swrlb:greaterThanOrEqual(?par, 300)  ∧
Coral_Reef:hasLongitude(?x, ?long)  ∧
swrlb:greaterThanOrEqual(?long, 147.5)  ∧
Coral_Reef:hasChemical_pH_levelOf(?x, ?ph)  ∧
swrlb:greaterThanOrEqual(?ph, 40)  ∧
Coral_Reef:hasChemicalSalinityLevelOf(?x, ?sal)  ∧
swrlb:greaterThanOrEqual(?sal, 50)
→ Coral_Reef:Bleached_Coral_Reef(?x)
```